Remarks

Claim Rejections 35 USC § § 101, 112 First Paragraph--Utility

The Examiner has rejected claims 1-35, 77 and 80-85 under 35 USC § 101 alleging that they are drawn to an invention with no apparent or disclosed patentable utility. First, the Examiner asserts that the polypeptide has been assigned a function because of its similarity to known proteins and then alleges that "it is commonly known in the art that sequence-to-function methods of assigning protein function are prone to error" citing Doerks, et al 1998, Brenner 1999 and Bork et al 1996.

The Examiner then goes on to state that even if *arguendo*, the nucleic acid encoding nGPCR-54 is found to be a G protein coupled receptor "its function is unknown" and that

> "Until some actual and specific significance can be attributed to the protein identified in the specification as nGPCR-54, the instant invention is incomplete. The polypeptide encoded by the nucleic acids of the invention is known to be structurally analogous to proteins that are known in the art as G protein coupled receptors. In the absence of knowledge of the natural substrate or biological significance of this protein, there is no immediately obvious patentable use for it."

The Applicants disagree.

I. The Applicable Legal Standard

To meet the utility requirement of sections 101 (and 112) of the Patent Act, the patent applicant need only show that the claimed invention is "practically useful," Anderson v. Natta, 480 F.2d 1392, 1397, 178 USPQ 458 (CCPA 1973) and confers a "specific benefit" on the public. Brenner v. Manson, 383 U.S. 519, 534-35, 148 USPQ 689 (1966). As discussed in a recent Court of Appeals for the Federal Circuit case, this threshold is not high:

> An invention is "useful" under section 101 if it is capable of providing some identifiable benefit. See Brenner v. Manson, 383 U.S. 519, 534 [148 USPQ 689] (1966); Brookiree Corp. v. Advanced Micro Devices, Inc., 977 F.2d 1555, 1571 [24 USPQ2d 1401] (Fed. Cir. 1992) ("to violate Section 101 the claimed device must be totally incapable of achieving a useful result"); Fuller v. Berger, 120 F. 274, 275 (7th Cir. 1903) (test for utility is whether invention "is incapable of serving any beneficial end"). Juicy Whip Inc. v. Orange Bang Inc., 51 USPQ2d 1700 (Fed. Cir. 1999).

While an asserted utility must be described with specificity, the patent applicant need not demonstrate utility to a certainty. In Stiftung v. Renishaw PLC, 945 F.2d 1173, 1180, 20 USPQ2d 1094 (Fed. Cir. 1991), the United States Court of Appeals for the Federal Circuit explained:

> An invention need not be the best or only way to accomplish a certain result, and it need only be useful to some extent and in certain applications: "[T]he fact that an invention has only limited utility and is only operable in certain applications is not grounds for finding lack of utility." Envirotech Corp. v. Al George, Inc., 730 F.2d 753, 762, 221 USPQ 473, 480 (Fed. Cir. 1984).

The specificity requirement is not, therefore, an onerous one. If the asserted utility is described so that a person of ordinary skill in the art would understand how to use the claimed invention, it is sufficiently specific. See Standard Oil Co. v. Montedison, S.p.a., 212 U.S.P.Q. 327, 343 (3d Cir. 1981). The specificity requirement is met unless the asserted utility amounts to a "nebulous expression" such as "biological activity" or "biological properties" that does not convey meaningful information about the utility of what is being claimed. Cross v. Lizuka, 753 F.2d 1040, 1048 (Fed. Cir. 1985).

In addition to conferring a specific benefit on the public, the benefit must also be "substantial." Brenner, 383 U.S. at 534. A "substantial" utility is a practical, "real-world" utility. Nelson v. Bowler, 626 F.2d 853, 856,206 USPQ 881 (CCPA 1980). As demonstrated in the Juicy Whip and Brooktree cases, supra, a mere "identifiable" benefit is substantial. So long as the claimed invention is not totally incapable of achieving a useful result, it meets the "substantiality" requirement. Ids.

If persons of ordinary skill in the art would understand that there is a "well-established" utility for the claimed invention, the threshold is met automatically and the applicant need not make any separate showing to demonstrate utility, regardless of what is disclosed in the patent specification. Manual of Patent Examination Procedure at § 706.03(a). Only if there is no "well-established" utility for the claimed invention must the applicant demonstrate the practical benefits of the invention. Id.

Once the patent applicant identifies a specific utility, the claimed invention is presumed to possess it. In re Cortright, 165 F.3d 1353, 1357, 49 USPQ2d 1464 (Fed. Cir. 1999); In re Brana,

51 F.3d 1560, 1566; 34 USPQ2d 1436 (Fed. Cir. 1995). In that case, the Patent Office bears the burden of demonstrating that a person of ordinary skill in the art would reasonably doubt that the asserted utility could be achieved by the claimed invention. Id. To do so, the Patent Office must provide evidence or sound scientific reasoning. See In re Langer, 503 F.2d 1380, 1391-92, 183 USPQ 288 (CCPA 1974). If and only if the Patent Office makes such a showing, the burden shifts to the applicant to provide rebuttal evidence that would convince the person of ordinary skill that there is sufficient proof of utility. Brana, 51 F.3d at 1566. The applicant need only prove a "substantial likelihood" of utility; certainty is not required. Brenner, 383 U.S. at 532.

II. The Examiner has misstated the Art as to Predictability of Associating Sequence with Function

As noted above the Examiner cites literature identifying some of the difficulties that may be involved in predicting protein function, none of the cited references suggests that functional homology cannot be inferred by a reasonable probability in any particular case. It is well-known that the probability that two unrelated polypeptides share more than 40% sequence homology over 70 amino acid residues is exceedingly small. Brenner et al., Proc. Natl. Acad. Sci. 95:6073-78 (1998) (**Exhibit 1**). Given homology in excess of 40% over many more than 70 amino acid residues, the probability that the polypeptide encoded for by our claimed polynucleotides is related to the reference polypeptides is, accordingly, very high. None of the Examiner's cited references contradicts Brenner's basic rule. Nor do they contradict our additional evidence of similarity to the G-protein coupled receptors, e.g., with respect to the presence of clearly delineated 7 transmembrane domains and conserved cysteine residues in the extracellular loops. At most, these articles cited by the USPTO individually and together stand for the proposition that it is difficult to make predictions about function with certainty. The standard applicable in this case is not, however, proof to certainty, but rather proof to reasonable probability as noted in the section above.

Under the Patent Law, the USPTO must accept the applicant's demonstration that the polypeptide encoded by the claimed invention is a member of a particular protein family and that utility is proven by a reasonable probability unless the USPTO can demonstrate through evidence or sound scientific reasoning that a person of ordinary skill in the art would doubt the asserted

utility. See In re Langer, 503 F.2d 1380, 1391-92, 183 USPQ 288 (CCPA 1974). The Examiner has simply not provided sufficient evidence or sound scientific reasoning to the contrary.

The Rejection Under 35 U.S.C. §101 Should be Withdrawn.

The Examiner has rejected claims 1-35, 77 and 80-85 alleging that the claimed invention is not supported by a specific and substantial asserted utility or a well-established utility. The Applicants respectfully traverse this rejection.

A. GPCR proteins have a well established utility.

Many medically significant biological processes are mediated by signal transduction pathways involving G-proteins and other second messengers, and G protein coupled seven transmembrane receptor proteins are recognized as important therapeutic targets for a wide range of diseases. According to a recently issued United States patent, nearly 350 therapeutic agents targeting GPCRs have been successfully introduced onto the market in only the last fifteen years. **(See U.S. Patent No. 6,114,127, at col. 2, lines 45-50.)** A recent journal review reported that most GPCR ligands are small and can be mimicked or blocked with synthetic analogues. That, together with the knowledge that numerous GPCRs are targets of important drugs in use today, make identification of GPCRs "a task of prime importance." **(See Exhibit 2**, Marchese et al., Trends Pharmacol. Sci., 20(9): 370-5., 1999.) Thus, the allegations that there is no well established utility for proteins ~~of the class that~~ the Applicants are now claiming is directly refuted by industry evidence. In this respect, the G protein coupled receptor family is analogous to the chemical genus that was the subject of In re Folkers, 145 USPQ 390 (CCPA 1965) (Compound that belongs to class of compounds, members of which are recognized as useful, is considered useful under §101.) The Patent Office does not serve the public by attempting to substitute a formulaic analysis of § 101 for the established judgment of the biopharmaceutical industry as to what is "useful." If the Patent Office is aware of any literature from the industry suggesting that GPCR's are not useful, the Applicants request that it be made of record.

Applicants would note for the record that the patent office apparently agrees with Applicant's reasoning in that it has granted and apparently continues to grant patents to G-protein coupled receptors, their encoding polynucleotides and antibodies directed to them in which no natural substrate/ligand or specific biological significance is ascribed to the protein. Specifically, Applicants would like to bring the following US Patents to the Examiner's attention:

US Patent 6,518,414 Maclennan "Molecular Cloning and Expression of G-Protein Coupled Receptors" (Claims an isolated polynucleotide)
US Patent 6,511,826 Li et al. "Polynucleotides Encoding Human G-Protein Chemokine Receptor (CCR5) HDGNR10" (Claims an isolated polynucleotide encoding a protein identified as a "chemokine receptor" with no specific chemokine identified)
US Patent 6,372,891 Soppet et al. "Human G-Protein Receptor HPRAJ70" (Claims an antibody directed to a G-protein coupled receptor)
US Patent 6,361,967 Agarwal et al. "AXOR10, A G-Protein Coupled Receptor" (Claims an isolated polynucleotide)
US Patent 6,348,574 Godiska et al. "Seven Transmembrane Receptors" (Claims an antibody directed to a G-protein coupled receptor)
US Patent 6,114,139 Hinuma et al. "G-Protein Coupled Receptor Protein and A DNA Encoding the Receptor" (Claims an isolated polynucleotide) Describe below.
US Patent 6,111,076 Fukusumi et al. "Human G-Protein Coupled Receptor (HIBCD07)" (Claims isolated polypeptide)
US Patent 6,107,475 Godiska et al. "Seven Transmembrane Receptors" (Claims isolated polynucleotide and methods)
US Patent 6, 096,868 Halsey et al. "ECR 673: A 7-Transmembrane G-Protein Coupled Receptor" (Claims isolated polypeptide)
US Patent 6,090,575 Li et al. "Polynucleotides Encoding Human G-Protein Coupled Receptor GPR1" (Claims isolated polynucleotide)
US Patent 6,071,722 Elshourbagy et al. "Nucleic Acids Encoding A G-Protein Coupled 7TM Receptor (AXOR-1)" (Claims an isolated polynucleotide)
US Patent 6, 071,719 Halsey et al. "DNA Encoding ECR 673: A 7-Transmembrane G-Protein Coupled Receptor" (Claims an isolated polynucleotide)
US Patent 6,060,272 Li et al. "Human G-Protein Coupled Receptors" (Claims isolated polynucleotide)
US Patent 6,048,711 Hinuma et al. "Human G-Protein Coupled Receptor Polynucleotides" (Claims isolated polynucleotide)
US Patent 6,030,804 Soppet et al. "Polynucleotides Encoding G-Protein Parathyroid Hormone Receptor HLTDG74 Polypeptides" (Claims isolated polynucleotide)
US Patent 6,025,154 Li et al. "Polynucleotides Encoding Human G-Protein Chemokine Receptor HDGNR10" (Claims an isolated polynucleotide encoding a protein identified as a "chemokine receptor" with no specific chemokine identified)
US Patent 5,998,164 Li et al. "Polynucleotides Encoding Human G-Protein Coupled Receptor GPRZ" (Claims isolated polynucleotide)
US Patent 5,994,097 Lal et al. "Polynucleotide Encoding Human G-Protein Coupled Receptor" (Claims isolated polynucleotide)
US Patent 5,958,729 Soppet et al. "Human G-Protein Receptor HCEGH45" (Claims isolated polypeptide)
US Patent 5,955,309 Ellis et al. "Polynucleotide Encoding G-Protein Coupled Receptor (H7TBA62)" (Claims isolated polynucleotide)
US Patent 5,948,890 Soppet et al. "Human G-Protein Receptor HPRAJ70" (Claims isolated polypeptide)

US Patent 5, 945,307 Glucksmann et al. "Isolated Nucleic Acid Molecules Encoding A G-Protein Coupled Receptor Showing Homology to The 5HT Family of Receptors" (Claims isolated polynucleotide)
US Patent 5, 942,414 Li et al. Polynucleotides Encoding Human G-Protein Coupled Receptor HIBEF51" (Claims isolated polynucleotide)
US Patent 5, 912,335 Bergsma et al. "G-Protein Coupled Receptor HUVCT36" (Claims isolated polynucleotide)
US Patent 5,874,245 Fukusumi et al. "Human G-Protein Coupled Receptors (HIBCD07)" (Claims isolated polynucleotide)
US Patent 5,871,967 Shabon et al. "Cloning of A Novel G-Protein Coupled 7TM Receptor" (Claims isolated polynucleotide)
US Patent 5,869,632 Soppet et al. "Human G-Protein Receptor HCEGH45" (Claims isolated polynucleotide)
US Patent 5,856,443 MacLennan et al. "Molecular Cloning and Expression of G-Protein Coupled Receptors" (Claims isolated polynucleotide)
US Patent 5,834,587 Chan et al. "G-Protein Coupled Receptor, HLTEX11" (Claims isolated polypeptide)
US Patent 5,776,729 Soppet et al. "Human G-Protein Receptor HGBER32" (Claims isolated polynucleotide)
US Patent 5,763,218 Fujii et al. "Nucleic Acid Encoding Novel Human G-Protein Coupled Receptors" (Claims isolated polynucleotide)
US Patent 5,756, 309 Soppet et al. "Nucleic Acid Encoding A Human G-Protein Receptor HPRAJ70 and Method of Producing the Receptor" (Claims isolated polynucleotide)
US Patent 5,585,476 MacLennan "Molecular Cloning and Expression of G-Protein Coupled Receptors" (Claims isolated polynucleotide)
US Patent 5,759,804 Godiska et al. "Isolated Nucleic Acid Encoding Seven Transmembrane Receptors" (Claims isolated polynucleotide and methods)

Applicants would submit these issued US Patents are evidence of an art recognized utility for G-protein coupled receptors whose natural ligand is unknown. If the Patent Office would take the position that issued patents are not sufficient evidence of art recognition then Applicants respectfully request that this position be made of record. In the alternative, if the Patent Office wishes to take the position that these issued patents are directed to non-statutory subject matter, then Applicants respectfully request that this position be made of record as well.

Furthermore, the Patent Office has neglected in its blanket statement:

"In the absence of knowledge of the natural substrate or biological significance of this protein, there is no immediately obvious patentable use for [a nucleic acid encoding] it ."

that it was well known in the art at the time of Applicant's filing that nucleic acids useful for measuring the expression of whole classes of genes are routinely incorporated for use in toxicology testing. Nuwaysir et al. *Microarrays and Toxicology:The Advent of Toxicogenomics* 24 Molecular Carcnogenesis 153 (1999)(**see Exhibit 3**) describes, for example, a Human ToxChip comprising 2089 human clones. The more genes that are available for use in toxicology testing, the more powerful the technique. "Arrays are at their most powerful when they contain the entire genome of the species they are being used to study." John C. Rockett and David J. Dix, *Application of DNA Arrays to Toxicology*, 107 Environ. Health Perspec. 681, No. 8 (1999)(**see Exhibit 4**). Control genes are carefully selected for their stability across a large set of array experiments in order to best study the effect of toxicological compounds. Thus, it is art recognized there is no expressed gene that is irrelevant to screening for toxicological effects, and all expressed genes have a utility for toxicological studies.

Even if the patent office were to ignore all of the aforementioned evidence of art recognition presented above, the historical success in the industry at developing therapeutics targeted to GPCR proteins supports a conclusion that the specific and substantial utilities for nGPCR-54, discussed in the next sections below, are entirely credible.

B. Screening for Ligands of nGPCR-54 Taught in the Application is a Specific and Substantial Utility

The use of the nGPCR-54 polypeptide (SEQ ID NO: 86) to screen for ligands that activate or inhibit nGPCR-54 is a specific and substantial utility. The use of a particular receptor such as nGPCR-54 to identify materials which specifically bind to that receptor is a specific utility because the method is not applicable to the general class of receptors. The method (which uses nGPCR-54 as a reagent) only identifies binding compounds for nGPCR-54, and cannot be expected to identify compounds that bind any other receptor. Stated differently, the identification of ligands which specifically bind to nGPCR-54 cannot be carried out with any integral membrane protein as asserted by the Examiner, but rather can only be carried out with nGPCR-54, if one hopes to have any reasonable expectation of success. The family of GPCRs is large, and the use of any other GPCR would not be expected to identify a ligand for nGPCR-54. Thus, a "specific" utility exists for nGPCR-54 polypeptides.

The identification of ligands which specifically bind nGPCR-54 is a substantial utility. In fact, the specification discloses neurological diseases in which nGPCR-54 may have utility due to inferred nGPCR-54 involvement in neurological functioning. (See Example 4, especially at p. 121, lines 30-35.) As explained in part A, above, the reported track record of successes in the pharmaceutical industry at targeting GPCR's (almost 350 marketed therapeutics in 15 years) supports a conclusion that such utility is substantial and credible.

C. The Use of nGPCR-54 as a Tissue Specific Probe is a Specific and Substantial Utility

nGPCR-54 tissue expression is detailed in Example 4, page 121 and brain specific expression is explored more fully at Example 11, page 146. Tissue specific markers are instrumental in pathology and other fields. If the Patent Office is aware of any scientific literature that has cast doubt of the practical utility of tissue specific markers, the Applicants request that it be made of record before issuance of any final action, to permit rebuttal and consideration on appeal.

D. The Use of nGPCR-54 as a Chromosomal Specific Probe is a Specific and Substantial Utility

At Example 13, page 148 Applicants localize the nGPCR-54 to chromosome 13 at position 13q32. Nucleic acids encoding nGPCR-54 and fragments thereof represent a reagent for whole chromosomal identification and identification of translocations of the relevant portion of chromosome 13. If the Patent Office is aware of any scientific literature that has cast doubt of the practical utility of chromosomal specific markers, the Applicants request that it be made of record before issuance of any final action, to permit rebuttal and consideration on appeal.

E. Conclusion

In conclusion, the use of nGPCR-54 polypeptides to identify ligands that specifically bind to nGPCR-54 has a specific and substantial utility in connection with neurological disorders.

The expression of nGPCR-54 in well defined tissues provides specific and substantial utility as a tissue marker. The expression of nGPCR-54 on chromosome 13 provides specific and substantial utility as a chromosomal marker. For these reasons, the rejection under §101 should be withdrawn.

III. The Rejection Under 35 U.S.C. §112, First Paragraph for Lack of Utility Should be Withdrawn.

In the Office Action, the Patent Office rejected claims 1-35, 77 and 80-85 alleging that the claimed invention is not supported by a specific and substantial or a well established utility. In support of the rejection, the Examiner relied on the utility rejection "set forth above." The Applicants respectfully traverse this rejection, for the reasons set forth above related to the utility rejection.

IV. The Rejection Under 35 U.S.C. §112, First Paragraph for Lack of Enablement Should be Withdrawn.

Claim 1 has been amended to make the Examiner's rejection moot. The term "homologous" has been deleted. The specification contains support for the amendment of claim 1 to include "fragments encoding a polypeptide comprising an epitope specific to said seven transmembrane receptor polypeptide" at page 29, lines 13-15 and at page 35 line 33 to page 36 through line 14. Claims 77 and 80-85 have been cancelled to facilitate prosecution so as to render the Examiner's rejection moot. Applicants reserve the right to present these claims in a later filed application.

Respectfully submitted,

Edward F. Rehberg, Attorney
Registration No. 34,703

Date: 5-2-2007

Pharmacia & Upjohn Company
Global Intellectual Property
301 Henrietta Street
Kalamazoo, Michigan 49001

Telephone No. (269) 833-7829 or (269) 833-9500
Telefax No. (269) 833-8897 or (269) 833-2316

Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 6073–6078, May 1998
Biochemistry

Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships

STEVEN E. BRENNER*†‡, CYRUS CHOTHIA*, AND TIM J. P. HUBBARD§

*MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, United Kingdom; and §Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambs CB10 1SA, United Kingdom

Communicated by David R. Davies, National Institute of Diabetes, Bethesda, MD, March 16, 1998 (received for review November 12, 1997)

ABSTRACT **Pairwise sequence comparison methods have been assessed using proteins whose relationships are known reliably from their structures and functions, as described in the SCOP database [Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia C. (1995) *J. Mol. Biol.* 247, 536–540]. The evaluation tested the programs BLAST [Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). *J. Mol. Biol.* 215, 403–410], WU-BLAST2 [Altschul, S. F. & Gish, W. (1996) *Methods Enzymol.* 266, 460–480], FASTA [Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* 85, 2444–2448], and SSEARCH [Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* 147, 195–197] and their scoring schemes. The error rate of all algorithms is greatly reduced by using statistical scores to evaluate matches rather than percentage identity or raw scores. The E-value statistical scores of SSEARCH and FASTA are reliable: the number of false positives found in our tests agrees well with the scores reported. However, the P-values reported by BLAST and WU-BLAST2 exaggerate significance by orders of magnitude. SSEARCH, FASTA ktup = 1, and WU-BLAST2 perform best, and they are capable of detecting almost all relationships between proteins whose sequence identities are >30%. For more distantly related proteins, they do much less well; only one-half of the relationships between proteins with 20–30% identity are found. Because many homologs have low sequence similarity, most distant relationships cannot be detected by any pairwise comparison method; however, those which are identified may be used with confidence.**

Sequence database searching plays a role in virtually every branch of molecular biology and is crucial for interpreting the sequences issuing forth from genome projects. Given the method's central role, it is surprising that overall and relative capabilities of different procedures are largely unknown. It is difficult to verify algorithms on sample data because this requires large data sets of proteins whose evolutionary relationships are known unambiguously and independently of the methods being evaluated. However, nearly all known homologs have been identified by sequence analysis (the method to be tested). Also, it is generally very difficult to know, in the absence of structural data, whether two proteins that lack clear sequence similarity are unrelated. This has meant that although previous evaluations have helped improve sequence comparison, they have suffered from insufficient, imperfectly characterized, or artificial test data. Assessment also has been problematic because high quality database sequence searching attempts to have both sensitivity (detection of homologs) and specificity (rejection of unrelated proteins); however, these complementary goals are linked such that increasing one causes the other to be reduced.

Sequence comparison methodologies have evolved rapidly, so no previously published tests has evaluated modern versions of programs commonly used. For example, parameters in BLAST (1) have changed, and WU-BLAST2 (2)—which produces gapped alignments—has become available. The latest version of FASTA (3) previously tested was 1.6, but the current release (version 3.0) provides fundamentally different results in the form of statistical scoring.

The previous reports also have left gaps in our knowledge. For example, there has been no published assessment of thresholds for scoring schemes more sophisticated than percentage identity. Thus, the widely discussed statistical scoring measures have never actually been evaluated on large databases of real proteins. Moreover, the different scoring schemes commonly in use have not been compared.

Beyond these issues, there is a more fundamental question: in an absolute sense, how well does pairwise sequence comparison work? That is, what fraction of homologous proteins can be detected using modern database searching methods?

In this work, we attempt to answer these questions and to overcome both of the fundamental difficulties that have hindered assessment of sequence comparison methodologies. First, we use the set of distant evolutionary relationships in the SCOP: Structural Classification of Proteins database (4), which is derived from structural and functional characteristics (5). The SCOP database provides a uniquely reliable set of homologs, which are known independently of sequence comparison. Second, we use an assessment method that jointly measures both sensitivity and specificity. This method allows straightforward comparison of different sequence searching procedures. Further, it can be used to aid interpretation of real database searches and thus provide optimal and reliable results.

Previous Assessments of Sequence Comparison. Several previous studies have examined the relative performance of different sequence comparison methods. The most encompassing analyses have been by Pearson (6, 7), who compared the three most commonly used programs. Of these, the Smith–Waterman algorithm (8) implemented in SSEARCH (3) is the oldest and slowest but the most rigorous. Modern heuristics have provided BLAST (1) the speed and convenience to make it the most popular program. Intermediate between these two is FASTA (3), which may be run in two modes offering either greater speed (ktup = 2) or greater effectiveness (ktup = 1). Pearson also considered different parameters for each of these programs.

To test the methods, Pearson selected two representative proteins from each of 67 protein superfamilies defined by the PIR database (9). Each was used as a query to search the database, and the matched proteins were marked as being homologous or unrelated according to their membership of PIR

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: EPQ, errors per query.

†Present address: Department of Structural Biology, Stanford University, Fairchild Building D-109, Stanford, CA 94305-5126
‡To whom reprints requests should be addressed. e-mail: brenner@hyper.stanford.edu.

superfamilies. Pearson found that modern matrices and "ln-scaling" of raw scores improve results considerably. He also reported that the rigorous Smith–Waterman algorithm worked slightly better than FASTA, which was in turn more effective than BLAST.

Very large scale analyses of matrices have been performed (10), and Henikoff and Henikoff (11) also evaluated the effectiveness of BLAST and FASTA. Their test with BLAST considered the ability to detect homologs above a predetermined score but had no penalty for methods which also reported large numbers of spurious matches. The Henikoffs searched the SWISS-PROT database (12) and used PROSITE (13) to define homologous families. Their results showed that the BLOSUM62 matrix (14) performed markedly better than the extrapolated PAM-series matrices (15), which previously had been popular.

A crucial aspect of any assessment is the data that are used to test the ability of the program to find homologs. But in Pearson's and the Henikoffs' evaluations of sequence comparison, the correct results were effectively unknown. This is because the superfamilies in PIR and PROSITE are principally created by using the same sequence comparison methods which are being evaluated. Interdependency of data and methods creates a "chicken and egg" problem, and means for example, that new methods would be penalized for correctly identifying homologs missed by older programs. For instance, immunoglobulin variable and constant domains are clearly homologous, but PIR places them in different superfamilies. The problem is widespread: each superfamily in PIR 48.00 with a structural homolog is itself homologous to an average of 1.6 other PIR superfamilies (16).

To surmount these sorts of difficulties, Sander and Schneider (17) used protein structures to evaluate sequence comparison. Rather than comparing different sequence comparison algorithms, their work focused on determining a length-dependent threshold of percentage identity, above which all proteins would be of similar structure. A result of this analysis was the HSSP equation; it states that proteins with 25% identity over 80 residues will have similar structures, whereas shorter alignments require higher identity. (Other studies also have used structures (18–20), but these focused on a small number of model proteins and were principally oriented toward evaluating alignment accuracy rather than homology detection.)

A general solution to the problem of scoring comes from statistical measures (i.e., E-values and P-values) based on the extreme value distribution (21). Extreme value scoring was implemented analytically in the BLAST program using the Karlin and Altschul statistics (22, 23) and empirical approaches have been recently added to FASTA and SSEARCH. In addition to being heralded as a reliable means of recognizing significantly similar proteins (24, 25), the mathematical tractability of statistical scores "is a crucial feature of the BLAST algorithm" (1). The validity of this scoring procedure has been tested analytically and empirically (see ref. 2 and references in ref. 24). However, all large empirical tests used random sequences that may lack the subtle structure found within biological sequences (26, 27) and obviously do not contain any real homologs. Thus, although many researchers have suggested that statistical scores be used to rank matches (24, 25, 28), there have been no large rigorous experiments on biological data to determine the degree to which such rankings are superior.

A Database for Testing Homology Detection. Since the discovery that the structures of hemoglobin and myoglobin are very similar though their sequences are not (29), it has been apparent that comparing structures is a more powerful (if less convenient) way to recognize distant evolutionary relationships than comparing sequences. If two proteins show a high degree of similarity in their structural details and function, it is very probable that they have an evolutionary relationship though their sequence similarity may be low.

The recent growth of protein structure information combined with the comprehensive evolutionary classification in the SCOP database (4, 5) have allowed us to overcome previous limitations. With these data, we can evaluate the performance of sequence comparison methods on real protein sequences whose relationships are known confidently. The SCOP database uses structural information to recognize distant homologs, the large majority of which can be determined unambiguously. These superfamilies, such as the globins or the immunoglobulins, would be recognized as related by the vast majority of the biological community despite the lack of high sequence similarity.

From SCOP, we extracted the sequences of domains of proteins in the Protein Data Bank (PDB) (30) and created two databases. One (PDB90D-B) has domains, which were all <90% identical to any other, whereas (PDB40D-B) had those <40% identical. The databases were created by first sorting all protein domains in SCOP by their quality and making a list. The highest quality domain was selected for inclusion in the database and removed from the list. Also removed from the list (and discarded) were all other domains above the threshold level of identity to the selected domain. This process was repeated until the list was empty. The PDB40D-B database contains 1,323 domains, which have 9,044 ordered pairs of distant relationships, or ≈0.5% of the total 1,749,006 ordered pairs. In PDB90D-B, the 2,079 domains have 53,988 relationships, representing 1.2% of all pairs. Low complexity regions of sequence can achieve spurious high scores, so these were masked in both databases by processing with the SEG program (27) using recommended parameters: 12 1.8 2.0. The databases used in this paper are available from http://sss.stanford.edu/sss/, and databases derived from the current version of SCOP may be found at http://scop.mrc-lmb.cam.ac.uk/scop/.

Analyses from both databases were generally consistent, but PDB40D-B focuses on distantly related proteins and reduces the heavy overrepresentation in the PDB of a small number of families (31, 32), whereas PDB90D-B (with more sequences) improves evaluations of statistics. Except where noted otherwise, the distant homolog results here are from PDB40D-B. Although the precise numbers reported here are specific to the structural domain databases used, we expect the trends to be general.

Assessment Data and Procedure. Our assessment of sequence comparison may be divided into four different major categories of tests. First, using just a single sequence comparison algorithm at a time, we evaluated the effectiveness of different scoring schemes. Second, we assessed the reliability of scoring procedures, including an evaluation of the validity of statistical scoring. Third, we compared sequence comparison algorithms (using the optimal scoring scheme) to determine their relative performance. Fourth, we examined the distribution of homologs and considered the power of pairwise sequence comparison to recognize them. All of the analyses used the databases of structurally identified homologs and a new assessment criterion.

The analyses tested BLAST (1), version 1.4.9MP, and WU-BLAST2 (2), version 2.0a13MP. Also assessed was the FASTA package, version 3.0t76 (3), which provided FASTA and the SSEARCH implementation of Smith–Waterman (8). For SSEARCH and FASTA, we used BLOSUM45 with gap penalties −12/−1 (7, 16). The default parameters and matrix (BLOSUM62) were used for BLAST and WU-BLAST2.

The "Coverage Vs. Error" Plot. To test a particular protocol (comprising a program and scoring scheme), each sequence from the database was used as a query to search the database. This yielded ordered pairs of query and target sequences with associated scores, which were sorted, on the basis of their scores, from best to worst. The ideal method would have

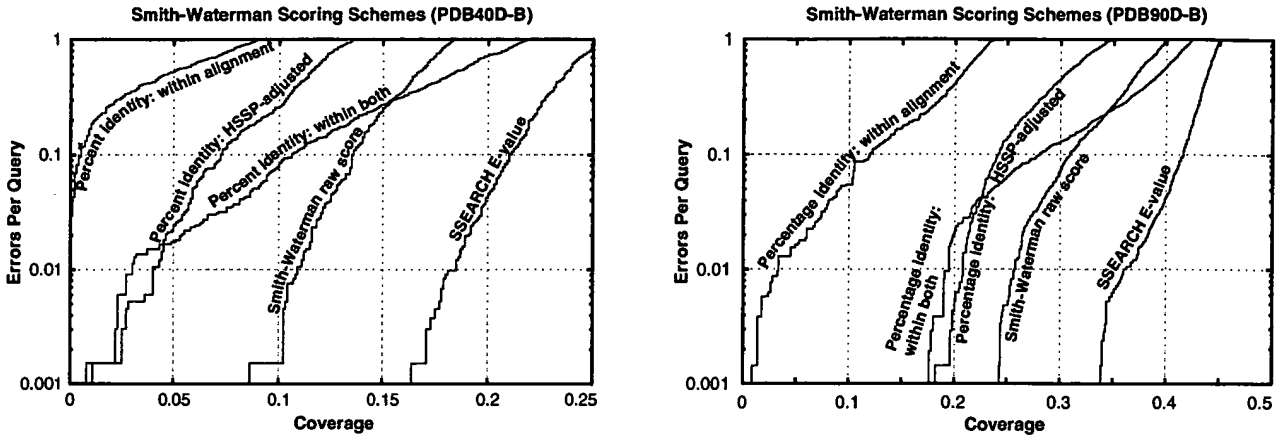

FIG. 1. Coverage vs. error plots of different scoring schemes for SSEARCH Smith–Waterman. (*A*) Analysis of PDB40D-B database. (*B*) Analysis of PDB90D-B database. All of the proteins in the database were compared with each other using the SSEARCH program. The results of this single set of comparisons were considered using five different scoring schemes and assessed. The graphs show the coverage and errors per query (EPQ) for statistical scores, raw scores, and three measures using percentage identity. In the coverage vs. error plot, the x axis indicates the fraction of all homologs in the database (known from structure) which have been detected. Precisely, it is the number of detected pairs of proteins with the same fold divided by the total number of pairs from a common superfamily. PDB40D-B contains a total of 9,044 homologs, so a score of 10% indicates identification of 904 relationships. The y axis reports the number of EPQ. Because there are 1,323 queries made in the PDB40D-B all-vs.-all comparison, 13 errors corresponds to 0.01, or 1% EPQ. The y axis is presented on a log scale to show results over the widely varying degrees of accuracy which may be desired. The scores that correspond to the levels of EPQ and coverage are shown in Fig. 4 and Table 1. The graph demonstrates the trade-off between sensitivity and selectivity. As more homologs are found (moving to the right), more errors are made (moving up). The ideal method would be in the lower right corner of the graph, which corresponds to identifying many evolutionary relationships without selecting unrelated proteins. Three measures of percentage identity are plotted. Percentage identity within alignment is the degree of identity within the aligned region of the proteins, without consideration of the alignment length. Percentage identity within both is the number of identical residues in the aligned region as a percentage of the average length of the query and target proteins. The HSSP equation (17) is $H = 290.15l^{-0.562}$ where l is length for $10 < l < 80$; $H > 100$ for $l < 10$; $H = 24.7$ for $l > 80$. The percentage identity HSSP-adjusted score is the percent identity within the alignment minus H. Smith–Waterman raw scores and E-values were taken directly from the sequence comparison program.

perfect separation, with all of the homologs at the top of the list and unrelated proteins below. In practice, perfect separation is impossible to achieve so instead one is interested in drawing a threshold above which there are the largest number of related pairs of sequences consistent with an acceptable error rate.

Our procedure involved measuring the coverage and error for every threshold. Coverage was defined as the fraction of structurally determined homologs that have scores above the selected threshold; this reflects the sensitivity of a method. Errors per query (EPQ), an indicator of selectivity, is the number of nonhomologous pairs above the threshold divided by the number of queries. Graphs of these data, called coverage vs. error plots, were devised to understand how protocols compare at different levels of accuracy. These graphs share effectively all of the beneficial features of Reciever Operating Characteristic (ROC) plots (33, 34) but better represent the high degrees of accuracy required in sequence comparison and the huge background of nonhomologs.

This assessment procedure is directly relevant to practical sequence database searching, for it provides precisely the information necessary to perform a reliable sequence database search. The EPQ measure places a premium on score consistency; that is, it requires scores to be comparable for different queries. Consistency is an aspect which has been largely

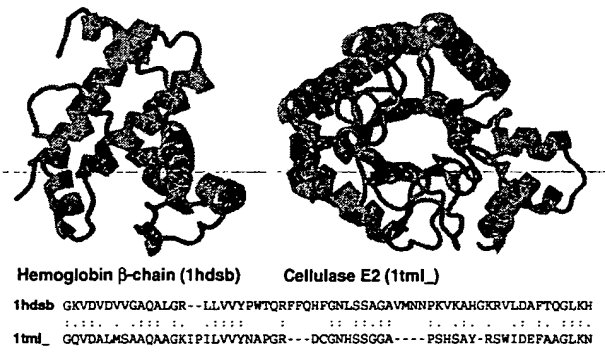

FIG. 2. Unrelated proteins with high percentage identity. Hemoglobin β-chain (PDB code 1hds chain b, ref. 38, *Left*) and cellulase E2 (PDB code 1tml, ref. 39, *Right*) have 39% identity over 64 residues, a level which is often believed to be indicative of homology. Despite this high degree of identity, their structures strongly suggest that these proteins are not related. Appropriately, neither the raw alignment score of 85 nor the E-value of 1.3 is significant. Proteins rendered by RASMOL (40).

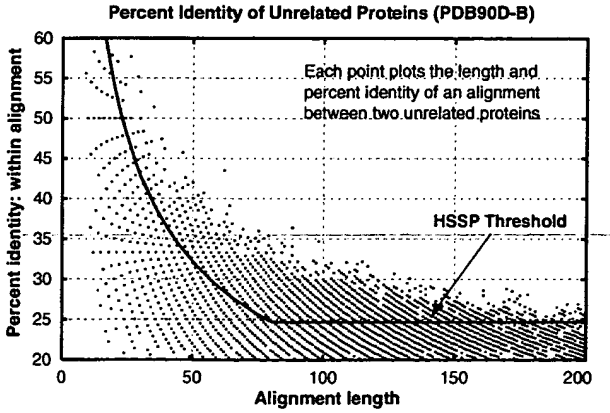

FIG. 3. Length and percentage identity of alignments of unrelated proteins in PDB90D-B: Each pair of nonhomologous proteins found with SSEARCH is plotted as a point whose position indicates the length and the percentage identity within the alignment. Because alignment length and percentage identity are quantized, many pairs of proteins may have exactly the same alignment length and percentage identity. The line shows the HSSP threshold (though it is intended to be applied with a different matrix and parameters).

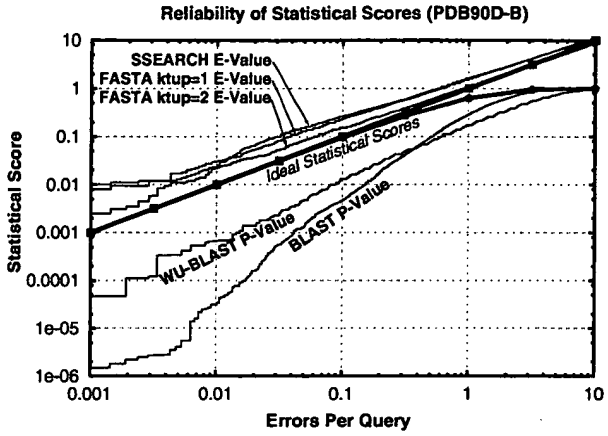

FIG. 4. Reliability of statistical scores in PDB90D-B: Each line shows the relationship between reported statistical score and actual error rate for a different program. E-values are reported for SSEARCH and FASTA, whereas P-values are shown for BLAST and WU-BLAST2. If the scoring were perfect, then the number of errors per query and the E-values would be the same, as indicated by the upper bold line. (P-values should be the same as EPQ for small numbers, and diverges at higher values, as indicated by the lower bold line.) E-values from SSEARCH and FASTA are shown to have good agreement with EPQ but underestimate the significance slightly. BLAST and WU-BLAST2 are overconfident, with the degree of exaggeration dependent upon the score. The results for PDB40D-B were similar to those for PDB90D-B despite the difference in number of homologs detected. This graph could be used to roughly calibrate the reliability of a given statistical score.

ignored in previous tests but is essential for the straightforward or automatic interpretation of sequence comparison results. Further, it provides a clear indication of the confidence that should be ascribed to each match. Indeed, the EPQ measure should approximate the expectation value reported by database searching programs, if the programs' estimates are accurate.

The Performance of Scoring Schemes. All of the programs tested could provide three fundamental types of scores. The first score is the percentage identity, which may be computed in several ways based on either the length of the alignment or the lengths of the sequences. The second is a "raw" or "Smith–Waterman" score, which is the measure optimized by the Smith–Waterman algorithm and is computed by summing the substitution matrix scores for each position in the alignment and subtracting gap penalties. In BLAST, a measure related to this score is scaled into bits. Third is a statistical score based on the extreme value distribution. These results are summarized in Fig. 1.

Sequence Identity. Though it has been long established that percentage identity is a poor measure (35), there is a common rule-of-thumb stating that 30% identity signifies homology. Moreover, publications have indicated that 25% identity can be used as a threshold (17, 36). We find that these thresholds, originally derived years ago, are not supported by present results. As databases have grown, so have the possibilities for chance alignments with high identity; thus, the reported cutoffs lead to frequent errors. Fig. 2 shows one of the many pairs of proteins with very different structures that nonetheless have high levels of identity over considerable aligned regions. Despite the high identity, the raw and the statistical scores for such incorrect matches are typically not significant. The principal reasons percentage identity does so poorly seem to be that it ignores information about gaps and about the conservative or radical nature of residue substitutions.

From the PDB90D-B analysis in Fig. 3, we learn that 30% identity is a reliable threshold for this database only for sequence alignments of at least 150 residues. Because one unrelated pair of proteins has 43.5% identity over 62 residues, it is probably necessary for alignments to be at least 70 residues in length before 40% is a reasonable threshold, for a database of this particular size and composition.

At a given reliability, scores based on percentage identity detect just a fraction of the distant homologs found by statistical scoring. If one measures the percentage identity in the aligned regions without consideration of alignment length, then a negligible number of distant homologs are detected. Use of the HSSP equation improves the value of percentage identity, but even this measure can find only 4% of all known homologs at 1% EPQ. In short, percentage identity discards most of the information measured in a sequence comparison.

Raw Scores. Smith–Waterman raw scores perform better than percentage identity (Fig. 1), but ln-scaling (7) provided no notable benefit in our analysis. It is necessary to be very precise when using either raw or bit scores because a 20% change in cutoff score could yield a tenfold difference in EPQ. However, it is difficult to choose appropriate thresholds because the reliability of a bit score depends on the lengths of the proteins matched and the size of the database. Raw score thresholds also are affected by matrix and gap parameters.

Statistical Scores. Statistical scores were introduced partly to overcome the problems that arise from raw scores. This scoring scheme provides the best discrimination between homologous proteins and those which are unrelated. Most

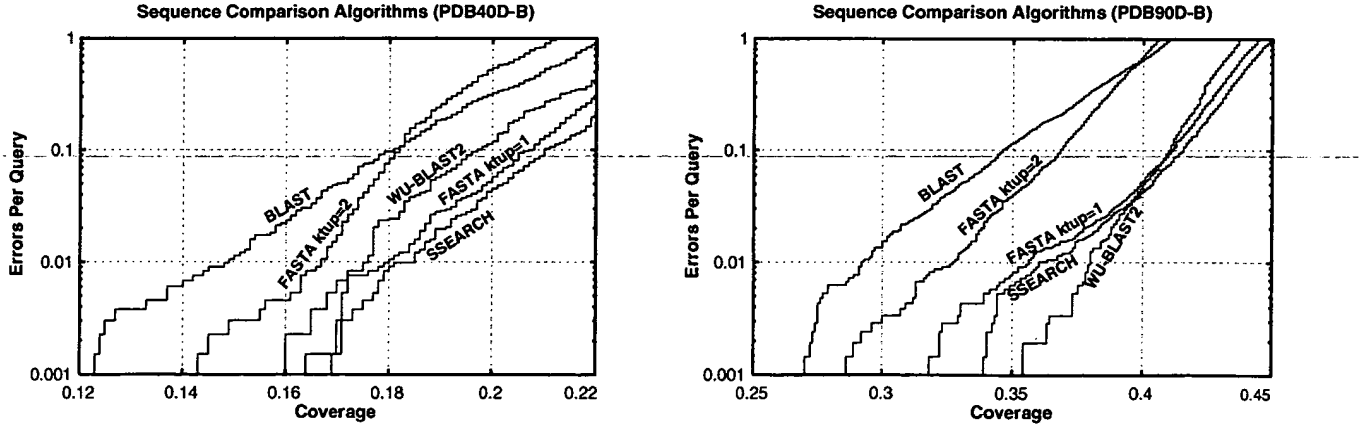

FIG. 5. Coverage vs. error plots of different sequence comparison methods: Five different sequence comparison methods are evaluated, each using statistical scores (E- or P-values). (*A*) PDB40D-B database. In this analysis, the best method is the slow SSEARCH, which finds 18% of relationships at 1% EPQ. FASTA ktup = 1 and WU-BLAST2 are almost as good. (*B*) PDB90D-B database. The quick WU-BLAST2 program provides the best coverage at 1% EPQ on this database, although at higher levels of error it becomes slightly worse than FASTA ktup = 1 and SSEARCH.

likely, its power can be attributed to its incorporation of more information than any other measure; it takes account of the full substitution and gap data (like raw scores) but also has details about the sequence lengths and composition and is scaled appropriately.

We find that statistical scores are not only powerful, but also easy to interpret. SSEARCH and FASTA show close agreement between statistical scores and actual number of errors per query (Fig. 4). The expectation value score gives a good, slightly conservative estimate of the chances of the two sequences being found at random in a given query. Thus, an E-value of 0.01 indicates that roughly one pair of nonhomologs of this similarity should be found in every 100 different queries. Neither raw scores nor percentage identity can be interpreted in this way, and these results validate the suitability of the extreme value distribution for describing the scores from a database search.

The P-values from BLAST also should be directly interpretable but were found to overstate significance by more than two orders of magnitude for 1% EPQ for this database. Nonetheless, these results strongly suggest that the analytic theory is fundamentally appropriate. WU-BLAST2 scores were more reliable than those from BLAST, but also exaggerate expected confidence by more than an order of magnitude at 1% EPQ.

Overall Detection of Homologs and Comparison of Algorithms. The results in Fig. 5*A* and Table 1 show that pairwise sequence comparison is capable of identifying only a small fraction of the homologous pairs of sequences in PDB40D-B. Even SSEARCH with E-values, the best protocol tested, could find only 18% of all relationships at a 1% EPQ. BLAST, which identifies 15%, was the worst performer, whereas FASTA ktup = 1 is nearly as effective as SSEARCH. FASTA ktup = 2 and WU-BLAST2 are intermediate in their ability to detect homologs. Comparison of different algorithms indicates that those capable of identifying more homologs are generally slower. SSEARCH is 25 times slower than BLAST and 6.5 times slower than FASTA ktup = 1. WU-BLAST2 is slightly faster than FASTA ktup = 2, but the latter has more interpretable scores.

In PDB90D-B, where there are many close relationships, the best method can identify only 38% of structurally known homologs (Fig. 5*B*). The method which finds that many relationships is WU-BLAST2. Consequently, we infer that the differences between FASTA kup = 1, SSEARCH, and WU-BLAST2 programs are unlikely to be significant when compared with variation in database composition and scoring reliability.

Fig. 6 helps to explain why most distant homologs cannot be found by sequence comparison: a great many such relationships have no more sequence identity than would be expected by chance. SSEARCH with E-values can recognize >90% of the homologous pairs with 30–40% identity. In this region, there are 30 pairs of homologous proteins that do not have significant E-values, but 26 of these involve sequences with <50 residues. Of sequences having 25–30% identity, 75% are identified by SSEARCH E-values. However, although the number of homologs grows at lower levels of identity, the detection falls off sharply: only 40% of homologs with 20–25% identity

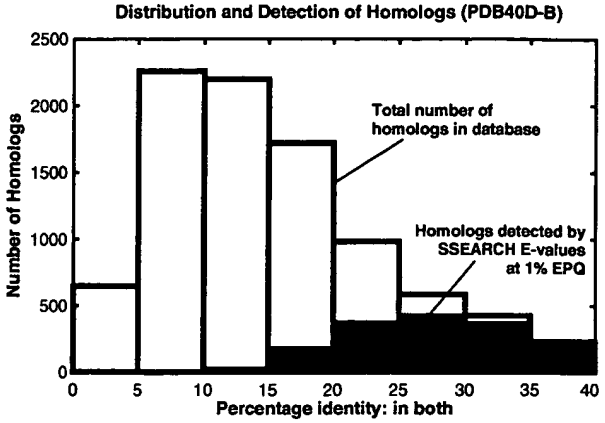

FIG. 6. Distribution and detection of homologs in PDB40D-B. Bars show the distribution of homologous pairs PDB40D-B according to their identity (using the measure of identity in both). Filled regions indicate the number of these pairs found by the best database searching method (SSEARCH with E-values) at 1% EPQ. The PDB40D-B database contains proteins with <40% identity, and as shown on this graph, most structurally identified homologs in the database have diverged extremely far in sequence and have <20% identity. Note that the alignments may be inaccurate, especially at low levels of identity. Filled regions show that SSEARCH can identify most relationships that have 25% or more identity, but its detection wanes sharply below 25%. Consequently, the great sequence divergence of most structurally identified evolutionary relationships effectively defeats the ability of pariwise sequence comparison to detect them.

are detected and only 10% of those with 15–20% can be found. These results show that statistical scores can find related proteins whose identity is remarkably low; however, the power of the method is restricted by the great divergence of many protein sequences.

After completion of this work, a new version of pairwise BLAST was released: BLASTGP (37). It supports gapped alignments, like WU-BLAST2, and dispenses with sum statistics. Our initial tests on BLASTGP using default parameters show that its E-values are reliable and that its overall detection of homologs was substantially better than that of ungapped BLAST, but not quite equal to that of WU-BLAST2.

CONCLUSION

The general consensus amongst experts (see refs. 7, 24, 25, 27 and references therein) suggests that the most effective sequence searches are made by (*i*) using a large current database in which the protein sequences have been complexity masked and (*ii*) using statistical scores to interpret the results. Our experiments fully support this view.

Our results also suggest two further points. First, the E-values reported by FASTA and SSEARCH give fairly accurate estimates of the significance of each match, but the P-values provided by BLAST and WU-BLAST2 underestimate the true

Table 1. Summary of sequence comparison methods with PDB40D-B

Method	Relative Time*	1% EPQ Cutoff	Coverage at 1% EPQ
SSEARCH % identity: within alignment	25.5	>70%	<0.1
SSEARCH % identity: within both	25.5	34%	3.0
SSEARCH % identity: HSSP-scaled	25.5	35% (HSSP + 9.8)	4.0
SSEARCH Smith–Waterman raw scores	25.5	142	10.5
SSEARCH E-values	25.5	0.03	18.4
FASTA ktup = 1 E-values	3.9	0.03	17.9
FASTA ktup = 2 E-values	1.4	0.03	16.7
WU-BLAST2 P-values	1.1	0.003	17.5
BLAST P-values	1.0	0.00016	14.8

*Times are from large database searches with genome proteins.

extent of errors. Second, SSEARCH, WU-BLAST2, and FASTA ktup = 1 perform best, though BLAST and FASTA ktup = 2 detect most of the relationships found by the best procedures and are appropriate for rapid initial searches.

The homologous proteins that are found by sequence comparison can be distinguished with high reliability from the huge number of unrelated pairs. However, even the best database searching procedures tested fail to find the large majority of distant evolutionary relationships at an acceptable error rate. Thus, if the procedures assessed here fail to find a reliable match, it does not imply that the sequence is unique; rather, it indicates that any relatives it might have are distant ones.**

**Additional and updated information about this work, including supplementary figures, may be found at http://sss.stanford.edu/sss/.

The authors are grateful to Drs. A. G. Murzin, M. Levitt, S. R. Eddy, and G. Mitchison for valuable discussion. S.E.B. was principally supported by a St. John's College (Cambridge, UK) Benefactors' Scholarship and by the American Friends of Cambridge University. S.E.B. dedicates his contribution to the memory of Rabbi Albert T. and Clara S. Bilgray.

1. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403–410.
2. Altschul, S. F. & Gish, W. (1996) *Methods Enzymol.* **266,** 460–480.
3. Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 2444–2448.
4. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995) *J. Mol. Biol.* **247,** 536–540.
5. Brenner, S. E., Chothia, C., Hubbard, T. J. P. & Murzin, A. G. (1996) *Methods Enzymol.* **266,** 635–643.
6. Pearson, W. R. (1991) *Genomics* **11,** 635–650.
7. Pearson, W. R. (1995) *Protein Sci.* **4,** 1145–1160.
8. Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* **147,** 195–197.
9. George, D. G., Hunt, L. T. & Barker, W. C. (1996) *Methods Enzymol.* **266,** 41–59.
10. Vogt, G., Etzold, T. & Argos, P. (1995) *J. Mol. Biol.* **249,** 816–831.
11. Henikoff, S. & Henikoff, J. G. (1993) *Proteins* **17,** 49–61.
12. Bairoch, A. & Apweiler, R. (1996) *Nucleic Acids Res.* **24,** 21–25.
13. Bairoch, A., Bucher, P. & Hofmann, K. (1996) *Nucleic Acids Res.* **24,** 189–196.
14. Henikoff, S. & Henikoff, J. G. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 10915–10919.
15. Dayhoff, M., Schwartz, R. M. & Orcutt, B. C. (1978) in *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M. (National Biomedical Research Foundation, Silver Spring, MD), Vol. 5, Suppl. 3, pp. 345–352.
16. Brenner, S. E. (1996) Ph.D. thesis. (University of Cambridge, UK).
17. Sander, C. & Schneider, R. (1991) *Proteins* **9,** 56–68.
18. Johnson, M. S. & Overington, J. P. (1993) *J. Mol. Biol.* **233,** 716–738.
19. Barton, G. J. & Sternberg, M. J. E. (1987) *Protein Eng.* **1,** 89–94.
20. Lesk, A. M., Levitt, M. & Chothia, C. (1986) *Protein Eng.* **1,** 77–78.
21. Arratia, R., Gordon, L. & M, W. (1986) *Ann. Stat.* **14,** 971–993.
22. Karlin, S. & Altschul, S. F. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 2264–2268.
23. Karlin, S. & Altschul, S. F. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 5873–5877.
24. Altschul, S. F., Boguski, M. S., Gish, W. & Wootton, J. C. (1994) *Nat. Genet.* **6,** 119–129.
25. Pearson, W. R. (1996) *Methods Enzymol.* **266,** 227–258.
26. Lipman, D. J., Wilbur, W. J., Smith, T. F. & Waterman, M. S. (1984) *Nucleic Acids Res.* **12,** 215–226.
27. Wootton, J. C. & Federhen, S. (1996) *Methods Enzymol.* **266,** 554–571.
28. Waterman, M. S. & Vingron, M. (1994) *Stat. Science* **9,** 367–381.
29. Perutz, M. F., Kendrew, J. C. & Watson, H. C. (1965) *J. Mol. Biol.* **13,** 669–678.
30. Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987) in *Crystallographic Databases: Information Content, Software Systems, Scientific Applications*, eds. Allen, F. H., Bergerhoff, G. & Sievers, R. (Data Comm. Intl. Union Crystallogr., Cambridge, UK), pp. 107–132.
31. Brenner, S. E., Chothia, C. & Hubbard, T. J. P. (1997) *Curr. Opin. Struct. Biol.* **7,** 369–376.
32. Orengo, C., Michie, A., Jones S, Jones D. T, Swindells M. B. & Thornton, J. (1997) *Structure (London)* **5,** 1093–1108.
33. Zweig, M. H. & Campbell, G. (1993) *Clin. Chem.* **39,** 561–577.
34. Gribskov, M. & Robinson, N. L. (1996) *Comput. Chem.* **20,** 25–33.
35. Fitch, W. M. (1966) *J. Mol. Biol.* **16,** 9–16.
36. Chung, S. Y. & Subbiah, S. (1996) *Structure (London)* **4,** 1123–1127.
37. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.
38. Girling, R., Schmidt, W., Jr, Houston, T., Amma, E. & Huisman, T. (1979) *J. Mol. Biol.* **131,** 417–433.
39. Spezio, M., Wilson, D. & Karplus, P. (1993) *Biochemistry* **32,** 9906–9916
40. Sayle, R. A. & Milner-White, E. J. (1995) *Trends Biochem. Sci.* **20,** 374–376.

64 Na, S. *et al.* (1996) *J. Biol. Chem.* 271, 11209–11213
65 Danley, D. E., Chuang, T-H. and Bokoch, G. M. (1996) *J. Immunol.* 157, 500–503
66 Mills, J. C., Stone, N. L., Erhardt, J. and Pittman, R. N. (1998) *J. Cell Biol.* 140, 627–636
67 Subauste, M. C. *et al. J. Biol. Chem.* (in press)
68 Billadeau, D. D. *et al.* (1998) *J. Exp. Med.* 188, 549–559
69 Rudel, T. and Bokoch, G. M. (1997) *Science* 276, 1571–1574
70 Cardone, M. H., Salvesen, G. S., Widmann, C., Johnson, G. and Frisch, S. M. (1997) *Cell* 90, 315–323
71 Chuang, T-H., Hahn, K. M., Lee, J-D., Danley, D. E. and Bokoch, G. M. (1997) *Mol. Biol. Cell* 8, 1687–1698
72 Lores, P., Morin, L., Luna, R. and Gacon, G. (1997) *Oncogene* 15, 601–605
73 Ward, C. *et al.* (1999) *J. Biol. Chem.* 274, 4309–4318
74 Sulciner, D. J. *et al.* (1996) *Mol. Cell. Biol.* 16, 7115–7121
75 Perona, R. *et al.* (1997) *Genes Dev.* 11, 463–475
76 Hirshberg, M., Stockley, R. W., Dodson, G. and Webb, M. R. (1997) *Nat. Struct. Biol.* 4, 147–152
77 Krengel, U. *et al.* (1990) *Cell* 62, 539–548
78 Ihara, K. *et al.* (1998) *J. Biol. Chem.* 273, 9656–9666
79 Abdul-Manan, N. *et al.* (1999) *Nature* 399, 379–383
80 Mott, H. R. *et al.* (1999) *Nature* 399, 384–388
81 Wu, W. J., Leonard, D. A., Cerione, R. A. and Manor, D. (1997) *J. Biol. Chem.* 272, 26153–26158
82 Fujisawa, K. *et al.* (1998) *J. Biol. Chem.* 273, 18943–18949
83 Scheffzek, K. *et al.* (1997) *Science* 277, 333–338
84 Scheffzek, K., Ahmadian, M. R. and Wittinghofer, A. (1998) *Trends Biochem. Sci.* 23, 257–262
85 Sprang, S. R. and Coleman, D. E. (1998) *Cell* 95, 155–158
86 Boriack-Sjodin, P. A., Margarit, S. M., Bar-Sagi, D. and Kuriyan, J. (1998) *Nature* 394, 337–343
87 Peyroche, A. *et al.* (1999) *Mol. Cell* 3, 275–285
88 Chardin, P. and McCormick, F. (1999) *Cell* 97, 153–155
89 Gibbs, J., Oliff, A. and Kohl, N. E. (1994) *Cell* 77, 175–178
90 Uehata, M. *et al.* (1997) *Nature* 389, 990–994
91 Kumar, C. C. *et al.* (1995) *Cancer Res.* 55, 5106–5117
92 Walsh, A. B., Dhanasekaran, M., Bar-Sagi, D. and Kumar, C. C. (1997) *Oncogene* 15, 2553–2560
93 Morozov, I., Lotan, O., Joseph, G., Gorzalczany, Y. and Pick, E. (1998) *J. Biol. Chem.* 273, 15435–15444
94 Kreck, M. L., Uhlinger, D. J., Tyagi, S. R., Inge, K. L. and Lambeth, J. D. (1994) *J. Biol. Chem.* 269, 4161–4168
95 Heyworth, P. G., Knaus, U. G., Settleman, J., Curnutte, J. T. and Bokoch, G. M. (1993) *Mol. Biol. Cell* 4, 1217–1223
96 Reif, K., Nobes, C. D., Thomas, G., Hall, A. and Cantrell, D. A. (1996) *Curr. Biol.* 6, 1445–1455
97 Zhou, K. *et al.* (1998) *J. Biol. Chem.* 273, 16782–16786
98 Manser, E. *et al.* (1998) *Mol. Cell* 1, 183–192

Acknowledgements
The authors were supported by grants from the National Institutes of Health (GM8), The National Arthritis Foundation, and the Association pour La Recherche sur le Cancer (VB). We wish to thank Dr Celine Der Mardirossian for thoughtful suggestions about the text and Antonetta Lestalla for secretarial assistance. Suggestions offered by the reviewers of this manuscript, which greatly improved the organization and content, are gratefully acknowledged.

Novel GPCRs and their endogenous ligands: expanding the boundaries of physiology and pharmacology

Adriano Marchese, Susan R. George, Lee F. Kolakowski Jr, Kevin R. Lynch and Brian F. O'Dowd

A. Marchese, Graduate Student, Dept of Pharmacology, Email: a.marchese@utoronto.ca
S. R. George*, Professor, Depts of Pharmacology and Medicine, Email: s.george@utoronto.ca
and **B. F. O'Dowd*,** Associate Professor, Dept of Pharmacology, University of Toronto, Medical Sciences Building, Toronto, ON, Canada M5S 1A8. Email: brian.odowd@utoronto.ca
*Also at Center for Addiction and Mental Health, 33 Russell Street, Toronto, ON, Canada M5S 2S1.
L. F. Kolakowski Jr, Associate Professor, Dept of Pharmacology, University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Drive, San Antonio, TX 78284-7763, USA. Email: kolakowski@uthscsa.edu
and **K. R. Lynch,** Professor, Dept of Pharmacology, University of Virginia Health Sciences Center, 1300 Jefferson Park Ave, Charlottesville, VA 22908, USA. Email: krl2r@virginia.edu

Nearly all molecules known to signal cells via G proteins have been assigned a cloned G-protein-coupled-receptor (GPCR) gene. This has been the result of a decade-long genetic search that has also identified some receptors for which ligands are unknown; these receptors are described as orphans (oGPCRs). More than 80 of these novel receptor systems have been identified and the emphasis has shifted to searching for novel signalling molecules. Thus, multiple neurotransmitter systems have eluded pharmacological detection by conventional means and the tremendous physiological implications and potential for these novel systems as targets for drug discovery remains unexploited. The discovery of all the GPCR genes in the genome and the identification of the unsolved receptor–transmitter systems, by determining the endogenous ligands, represents one of the most important tasks in modern pharmacology.

The G-protein-coupled receptors (GPCRs) are transducers of extracellular messages and they allow tissues to respond to a wide array of signalling molecules. Most of the endogenous ligands are small and the binding of these ligands to their receptor(s) can be mimicked (or blocked) by synthetic analogues. Together with the knowledge that numerous GPCRs are targets of important drugs in use today, GPCR identification is a task of prime importance. In the 14 years since the first cloning of genes for GPCRs, most of the molecules known to signal cells via the heterotrimeric G-protein–effector systems have been assigned a cloned GPCR gene. However, the vigorous search for novel GPCR genes has far outpaced the identification of novel endogenous ligands. A group of genes has been identified whose products are, using the criterion of sequence similarity, members of the GPCR family but for which the ligands are not known, and these are commonly known as orphans (oGPCR).

The GPCR gene family is the largest known receptor family (see Box 1) and shares a common secondary structure that consists of seven transmembrane domains. Setting aside the odorant receptors (encoded by hundreds of genes), nearly 300 mammalian GPCR genes have been recognized[1]. On the basis of structure, the GPCRs can be separated into three subfamilies. The inclusion of a receptor in a subfamily requires the presence of an overall percentage amino acid identity and not any discrete motif. Most GPCRs, including the odorant receptors, are grouped in Family A. Several additional GPCRs, which have as their ligands peptides such as secretin, vasoactive intestinal peptide and calcitonin, make up Family B. Family C comprises the metabotropic glutamate receptors, the Ca^{2+}-sensing receptor, pheromone receptors, the $GABA_B$ receptors and the taste receptors. Within each family, GPCRs are grouped by sequence similarity and ligand specificity; approximately one third of Family A members

Box 1. How big is the GPCR family?

The size f the GPCR family surprised even the most optimistic pharmacologist as many subfamilies proved to be larger than had been predicted by classical pharmacological techniques. Furthermore, some ligands that were not widely considered to signal via receptors (e.g. nucleotides) are recognized now to have numerous receptor subtypes. The discovery of these multiple subtypes, new ligands and the rapid accumulation of novel GPCR sequences have led to the expectation that many more mammalian GPCRs await discovery. Thus, an obvious question to ask is: how many GPCR genes are there in the human genome? Although simply waiting a few years should answer this question directly, there are practical implications in making an educated guess now. For example, is the receptor for a candidate ligand likely to be visible now among the existing oGPCR DNAs? And, is further searching for oGPCR DNAs a worthwhile endeavour?

The recent completion of the nematode (*Caenorhabditis elegans*) translated genome provides an interesting comparison to mammalian GPCRs. In contrast to the single cell yeast (with its two GPCR genes), multicellularity obviously demands cell-to-cell communication and the added complexity imposes a requirement for a much larger repertoire of GPCRs. According to th analysis reported by Bargmann[1], 5% of the 19100 nematode genes encode GPCRs. Their distribution among GPCR families is reminiscent of the mammalian GPCR genes, some 700–1000 chemoattractant (odorant) genes (including numerous pseudogenes), approximately 150 Family A genes and four-to-five each Family B and C genes. By analogy, this suggests that the number of mammalian GPCRs could total 5000 (5% of mammalian genes estimated to be 80 000–100 000). Unfortunately, the *C. elegans* genome provides no direct clues for oGPCR identification as the closest nematode GPCR is <35% identical to any mammalian GPCR and there are no obvious homologues to mammalian pre-pro-neuropeptide genes. In contrast, the accumulation of nucleotide sequence information from another surrogate organism, the zebrafish (*Danio rerio*), should be more informative because the conceptualized GPCR amino acid sequences are often ~70% identical to orthologous mammalian GPCRs.

Reference
1 Bargmann, C. (1998) *Science* 282, 2028–2033

are oGPCRs and this review will focus on these receptors. Thus, in a decade, the list of signalling molecules for which the GPCR genes had not been cloned has been supplanted by a list of ~80 oGPCRs awaiting a ligand (see Table 1). The characterization of these GPCRs has already enabled the discovery of several new endogenous ligands; this will be discussed later.

Novel GPCR gene discovery

Very few GPCRs have been purified, thus the pace of GPCR gene discovery has been fuelled by a series of highly successful cloning techniques. The identification (using amino acid sequence determination and expression cloning) of a few sequences encoding Family A GPCRs demonstrated that these were related genes[1]. Cloning by low stringency hybridization to cDNA/genomic DNA libraries yielded a stream of novel GPCR DNAs. The pace of discovery quickened with the use of the polymerase chain reaction (PCR). The database of expressed sequence tagged cDNAs (ESTs) has provided material for a further expansion of Family A, as has the high-throughput sequencing of 100–200 kb pair segments of human DNA.

Novel GPCR identification

Many oGPCRs are found to be similar to known GPCRs. Where the identity reaches the threshold of ~45%, it is likely that the receptors will share a common ligand, i.e. that the oGPCR will be a pharmacological subtype of the known GPCR. This rule is not without exception. Take, for example the orphanin FQ/nociceptin receptor; this has ~65% amino acid identity to opioid receptors, but does not have high affinity for opioid peptides[2,3]. Many GPCR subtypes have <40% amino acid identity, in which case sequence comparison might not be profitable. Moreover, because the ligand-binding pocket has not yet been described fully for any receptor, it is not feasible to predict ligand identity. However, dendritic tree building shows that receptors that respond to the same, or similar, agonists often cluster. For example, most members of the prostanoid receptor subfamily share <30% amino acid identity, yet these eight receptors are more like one another than any other GPCR. A similar situation exists among the nucleotide receptors, chemokine receptors and other cationic amine receptors. In the way that many known GPCRs fall into subfamilies, many oGPCRs cluster together, sometimes with members having >50% amino acid identity, which suggests that the problem of the ~80 oGPCRs might be solved by a mere 30 or 40 ligands. For example, the recent identification of Edg-1 as a sphingosine 1-phosphate receptor[4–6] leads directly to the prediction that Edg-3 and Edg-5 (both >50% identical to Edg-1) have the same ligand. More distant members of the Edg cluster, Edg-2 and Edg-4 are known to be receptors for the structurally related ligand, lysophosphatidic acid[7–9].

When homology does not inform, i.e. the nearest known GPCR has <35% amino acid identity to the orphan, ligand identification is challenging. There are no signature amino acids that predict either the nature of the ligand or the identity of the interacting Gα subunit type(s). In those cases where the ligand is a molecule with an established pharmacology, tissue distribution has allowed inference of ligand identity. Thus, an important clue to identifying the oGPCR RDC-8 as encoding the adenosine A_{2A} receptor was the concordance of *in situ* hybridization and ligand ([^{3}H]CGS21680) autoradiography signals in rat brain sections[10]. Similarly, the occurrence of both cannabinoid binding sites and SKR6 receptor mRNA accumulation in NG108 cells led to the identification of the cannabinoid CB_1 receptor[11].

Table 1. Amino acid sequence identity of some orphan G-protein-coupled receptors

Homology	Name	Species	% Amino acid identity	Accession no.
Opioid and somatostatin receptor-like	GPR7	Human	62% GPR8, 40% sst_5	U22491
	GPR8	Human	62% GPR7, 45% sst_5	U22492
	GPR24	Human	33% sst_2, 32% sst_5	U71092
	GPR14	Rat	29% μ-opioid, 28% sst_4	U32673
	GPR54	Rat	37% gal2, 35% GAL1	AF115516
Chemokine receptor-like	GPR2	Human	41% CXCR3, 40% CCR7	U13667
	CKRX	Human	53% EO1, 43% CCR1	AF014958
	EO1	Mouse	53% CKRX, 36% CCR1	AF030185
	MIP-1αRL1	Mouse	62% CCR1, 50% CCR3	U28405
	GPR28	Human	43% CCR7, 38% CCR6	U45982
	STRL33	Human	37% CCR7, 37% CCR6	U73529
	PPR1	Bovine	39% CCR7, 37% GPR28	S63848
	g10d	Rat	33% RDC1, 30% CCR9	L09249
	RDC1	Human	33% g10d, 30% CXCR2	X14048
	TM7SF1	Human	22% GPR5, 14% CCR6	AF027826
	CLR1	Chicken	51% BLR1, 36% CXCR1	AF029369
Chemoattractant receptor-like	Dez	Human	37% GPR1, 35% FPR2	U79527
	FPRL2	Human	72% FPR2, 56% FPR1	M76673
	FPR2	Human	72% FPRL2, 69% FPR1	M76672
	GPR1	Human	37% Dez, 34% FPR2	U13666
	GPR30	Human	32% FPRL2, 32% FPR2	AF027956
	GPR32	Human	39% FPR1, 35% FPRL2	AF045764
	GPR33	Mouse	36% GPR32, 36% Dez	AF045766
	GPR44	Human	37% Dez, 36% FPRL2	AF118265
	mas oncogene	Human	34% MRG, 26% C5aR	M13150
	MRG	Human	34% *mas* oncogene, 34% C5aR	S78653
	RTA	Rat	32% *mas* oncogene, 33% MRG	M32098
	GPR53p	Human	35% MRG, 28% *mas* oncogene	AF096785
Angiotensin receptor-like	GPR15	Human	34% GPR25, 31% APJ	U34806
	GPR25	Human	34% GPR15, 32% APJ	U91939
Cannabinoid receptor-like	GPR3	Human	59% GPR6, 57% GPR12	U13668
	GPR6	Human	59% GPR3, 56% GPR12	L36150
	GPR12	Rat	57% GPR3, 56% GPR6	U18548
	EDG-6	Human	46% EDG-3, 44% EDG-1	AJ000479
GPR4 receptor-like	OGR1	Human	48% GPR4, 35% TDAG8	U48405
	GPR4	Human	48% GPR12A, 36% TDAG8	L36148
	TDAG8	Human	36% GPR4, 35% GPR12A	U95218
	G2A	Mouse	34% GPR4, 31% OGR1	AF083442
Neuropeptide Y receptor-like	GIR	Mouse	35% GPR10, 30% NK_4	M80481
	GPR19	Human	27% GAL1, 26% NPY Y_2	U64871
	GPR22	Human	26% NPY Y_5, 24% CCK_1	U66581
Amine receptor-like	PNR	Human	33% 5-HT_4, 33% 5-HT_7	AF021818
	GPR26	Human	28% 5-HT_{5B}, 23% 5-HT_{5A}	
	GPR27	Mouse	29% D4, 25% 5-HT_6	AF027955
	AGR9	Rat	24% H_2, 24% NK_2	S73608
	GPR21	Human	27% β_3AR, 24% β_2AR	U66580
	PSP24	Human	26% 5-HT_4, 23% β_1AR	U92642
	GPR45	Human	70% PSP24, 21% NK_2	AF118266
	A-2	Human	21% 5-HT_{1F}, 19% 5-HT_{1E}	U47928
	GPR52	Human	71% GPR21, 27% H_2	AF096784
	RE2	Human	25% α_{1A}AR, 25% α_{1C}AR	AF091890
	GPR57	Human	59% GPR58, 37% PNR	N/A
	GPR58	Human	59% GPR57, 42% PNR	N/A
	GPR61	Human	27% LZY2, 30% 5-HT_6	N/A
	GPR62	Human	27% LZY, 28% 5-HT_6	N/A
P2 receptor-like	GPR23	Human	53% RBintron, 33% '$P2Y_{10}$'	U66578
	RBintron	Human	53% GPR23, 38% $P2Y_4$	L11910
	GPR35	Human	32% GPR23, 30% HM74	AF027957
	'$P2Y_{10}$'	Human	34% RBintron, 33% GPR23	AF000545
	GPR17	Human	35% $P2Y_2$, 34% $P2Y_4$	U33447
	GPR18	Human	30% RBintron, 29% GPR17	L42324
	HM74	Human	36% GPR31, 29% $P2Y_1$	D10923
	GPR31	Human	36% HM74, 29% $P2Y_1$	U65402

Table 1. (*cont.*)

Homology	Name	Species	% Amino acid identity	Accession no.
P2 receptor-like (*cont.*)	RSC338	Human	33% H963, 28% tp2y	D13626
	EBI 2	Human	33% RBintron, 30% CCR1	L08177
	H963	Human	33% RSC338, 28% PAFR	AF002986
	GPR41	Human	98% GPR42, 41% GPR43	AF024688
	GPR42	Human	98% GPR41, 28% GPR23	AF024689
	GPR40	Human	31% GPR43, 26% CXCR1	AF024687
	GPR43	Human	41% GPR41, 31% GPR40	AF024690
	GPR20	Human	31% $P2Y_4$, 26% GPR23	U66579
	GPR34	Human	31% RSC338, 29% RBintron	AF118670
	GPR55	Human	29% $P2Y_5$, 30% GPR23	AF096786
Neurotensin receptor-like	GHS-R	Human	35% NTS1, 33% nts2	U60179
	GPR39	Human	32% NTS1, 25% nts2	AF034633
	HSOGPCR2	Human	38% GPR38, 34% GHS-R	AF044601
Melatonin receptor-like	H9	Human	48% ML_{1A}, 45% ML_{1B}	U52219
Endothelin receptor-like	GPR37	Human	68% ET_BR-LP-2, 27% ET_B	U87460
	ETBR-LP-2	Human	68% GPR37, 27% ET_B	Y16280
Glycoprotein hormone receptor-like	LGR5	Human	26% FSH-R, 25% LH-R	AF062006
Opsin receptor-like	Encephalopsin	Human	32% Peropsin, 31% Rhodopsin	AF140242
	RGR	Human	27% Peropsin, 26% Rhodopsin	U15790

Please refer to the *TiPS Receptor and Ion Channel Nomenclature Supplement* and to individual GenBank accession numbers for further information.

Endogenous ligand identification

In the same way that EST database searching has yielded GPCR DNAs, it has also yielded DNAs encoding peptide sequences related to known peptides. Several novel chemokines have been discovered using this approach and these have proven to be the ligands for several chemokine receptors. For example, a CC chemokine termed ELC (EBI-ligand chemokine) was identified from the EST database and found to be the endogenous ligand for the orphan receptor EBI1, which has since been renamed CCR7 (Ref. 12). Similarly, the CC chemokine liver and activation-regulated chemokine (LARC) was identified from the EST database[13] and subsequently shown to be the ligand for the orphan STRL22 receptor; this was renamed CCR6 (Refs 14–16). Another EST encoding a CXC chemokine was isolated, BCA1 (Ref. 17), and later identified as a ligand for the oGPCR BLR1, which has since been renamed CXCR5 (Ref. 18). A fourth, novel class of chemokines called δ-chemokines, or CX_3C chemokines, was discovered by automated high-throughput single-pass sequencing and analysis of a cDNA library constructed from murine choroid plexus[19]. The sequence of one of the cDNA clones exhibited similarity to murine monocyte chemoattractant protein-1 (MCP-1), an α-chemokine. Also, another group independently searched the EST database with known chemokine sequences and identified the same chemokine, which they have termed fractalkine[20]. This ligand was matched to the orphan receptor V28 (renamed CX3CR1)[21]. The ligand for the novel receptor encoded by GPR5 (Ref. 22) has been identified as the single C motif-1 peptide[23] and the receptor renamed as XC chemokine receptor I. The ongoing search for the discovery of novel chemokines will most certainly reveal novel candidates to test with the existing chemokine-like orphan receptors and any additional genes encoding chemokine receptors.

With oGPCR DNAs in hand and with nearly all known ligands assigned, the task now is to use oGPCR DNAs to discover novel ligands[24]. The strategy employed is to express the oGPCR DNA in a cell and apply tissue extracts until a response is observed. The agonist ligand is then purified, synthesized and re-tested. This approach has been most successful in identifying neuropeptides. Peptide ligands often exhibit high-affinity interactions with their receptors, which enables detection at low concentrations and the development of radioligand binding assays. The first success at orphan ligand identification involved a GPCR with sequence identity to the opioid receptors. The natural ligand was identified by two research groups using brain extracts[2,3] and the peptide discovered was 17 amino acids in length, named either orphanin FQ or nociceptin. The peptide contains the tetrapeptide FGGF, which is similar to the motif YGGF of the opioid peptides. Another successful strategy used rat brain fractions that were applied to cells and Ca^{2+} mobilization measured; this succeeded in identifying a novel brain peptide. This peptide and a related peptide (from the same precursor protein) bound to two related oGPCRs and these peptides, which are found in the hypothalamus, function in appetite regulation and satiety control and thus were named orexins[25] (also known as hypocretins[26]). In a similar series of experiments, Hinuma *et al.*[27] measured arachidonate release from CHO cells transfected with the GPR10 (Ref. 28) to identify a novel brain peptide with prolactin-releasing properties at the anterior pituitary. This group has also identified another novel peptide, apelin[29], as the ligand for the receptor APJ (Ref. 30).

The elusive nature of certain labile natural agonists could be a significant hindrance to the discovery of oGPCR ligands, as there is no reason to believe that the remaining oGPCR ligands will all prove to be peptides. An attempt to address this problem involves the use of combinatorial chemistry to generate large libraries of compounds to be tested as surrogate agonists. Although not the physiological solution to the problem, such compounds are tools for probing the pharmacology of an oGPCR. Recently, an interesting variation to this approach was reported. Yeast expressing the human formyl peptide receptor-like oGPCR, FPR2 (Ref. 31), was made dependent on stimulation of this receptor for growth in histidine-free medium and then transfected with a plasmid DNA library designed to express random tridecapeptides. Yeast colonies that were no longer dependent on histidine were judged to have undergone autocrine stimulation and the responsible plasmids recovered. The results yielded a set of six peptides, one of which elicited Ca^{2+} mobilization in HEK293 cells transfected with the FPR2 plasmid.

Ligand-sceening assays

There has been a concerted effort to make ligand identification more efficient by developing cell-based assay systems that have low endogenous GPCR background or report G-protein activation events, or both, in a robust, readily detected manner. The existence of endogenous GPCR signalling systems is important because overexpression of one GPCR can elicit an exaggerated response via other, unrelated and previously unrecognized endogenous GPCRs (Ref. 32), and this could result in false positives. The aforementioned yeast expression system is attractive because of the absence of many endogenous GPCRs. In essence, it involves replacing the endogenous pheromone receptor with a mammalian GPCR and redirecting the pheromone pathway response from a mitogen-activated protein kinase type activation to a biosynthetic circuit, thus allowing the synthesis of histidine. In this case, agonist stimulation allows growth on histidine-free medium. Potential drawbacks of the yeast expression system are the difficulties in expressing some GPCRs achieving effective receptor–G-protein coupling and ligand binding to yeast cell wall components.

Another assay system, which uses mammalian cells, takes advantage of the relatively high expression levels achieved following transfection of oGPCR DNAs so that the endogenous, low-level receptors do not interfere. This system uses the translocation of β-arrestin to receptor sites on the plasma membrane after agonist-mediated receptor activation. Barak *et al.* have shown, using a β-arrestin-2/green fluorescent protein (βarr2–GFP) fusion protein and confocal microscopy, that on agonist stimulation of the β_2-adrenoceptor, βarr2–GFP translocates to the plasma membrane, and that this interaction can be enhanced by co-expression of G-protein-coupled receptor kinase 2 (Ref. 33). This group also showed that similar responses are observed with other receptors coupled to different G proteins, which suggests that the cellular visualization of the agonist-mediated translocation of βarr2–GFP could provide a widely applicable method for detecting the activation of GPCRs.

A system that is useful in measuring GPCR-mediated activation of $G\alpha_q$, $G\alpha_{i/o}$ and $G\alpha_s$ is based on pigment dispersion or aggregation in cultured *Xenopus laevis* melanophores[34,35]. Increases in cAMP ($G\alpha_s$-coupled receptors) or activation of protein kinase C ($G\alpha_q$) lead to pigment dispersion causing darkening of the cells, while decreases in cAMP ($G\alpha_{i/o}$) lead to pigment aggregation near the nucleus and make the cells appear clear[36]. These colour changes are detected readily, however these cells have a substantial complement of endogenous GPCRs, which could confound the results. Overexpression of receptors in melanophores results in changes in the 'basal' signalling and promotes either the clear or the dark cell colour, thus predicting either $G\alpha_{i/o}$ signalling or $G\alpha_q$ or $G\alpha_s$ pathways.

A simpler approach to detecting the activation of multiple types of G proteins uses Gα16 as a universal adapter G protein that can funnel the signal-transduction machinery down a common pathway, such that a single second-messenger response (Ca^{2+} mobilization) can be measured for a given receptor[37]. Heterologous expression of Gα16 allows the coupling of a wide range of GPCRs to phospholipase activity, and thence to Ca^{2+} mobilization. For example, the β_2-adrenoceptor normally couples only to $G\alpha_s$, but when the β_2-adrenoceptor and Gα16 are transiently co-expressed in COS7 cells agonist-dependent stimulation results in inositol phosphate (IP) production[38]. Receptors linked to $G\alpha_s$ (e.g. dopamine D1, vasopressin V_2 and adenosine A_{2A} receptors) or pertussis-toxin-sensitive $G\alpha_i$ (e.g. muscarinic acetylcholine M_2, 5-HT_{1A}, formyl-peptide FPR1 and δ-opioid receptors), when co-transfected with Gα16, also caused concentration-dependent, agonist-mediated IP generation[38]. Other receptors (e.g. thromboxane A_2 and vasopressin V_1) that routinely couple to $G\alpha_q$ and Gα11 to stimulate IP generation were also shown to couple effectively to Gα15 and Gα16 (Ref. 38). However, this coupling is not universal, as the chemokine receptor, CCR1, that effectively couples to $G\alpha_i$ and $G\alpha_q$, failed t couple to Gα16 (Ref. 39).

Other considerations

Recently, new complexities have been added to the general approach to studying orphan GPCRs. For instance, the oGPCR calcitonin receptor-like receptor, has been cloned[40]. The expression of this receptor was consistent with the expression pattern of a calcitonin gene-related peptide (CGRP). The efficient binding of CGRP or amylin, or both, to this receptor required the co-expression of a cofactor protein called receptor activity modifying protein 1 (RAMP1)[41].

Studies have shown that heterodimerization of two GPCR subunits are required for the formation of a functional $GABA_B$ receptor[42–46]. The apparent requirement for two different gene products to create a GPCR signalling entity indicates that the characterization of some oGPCRs might be more complex, perhaps indicating that functional

assays should begin to include co-expression of related oGPCRs.

In principle, the elimination of a GPCR gene from the germline and testing the resulting knockout mice for some change might provide clues to GPCR function, if not ligand identity. For example, when the mouse BLR1 orphan receptor was disrupted, it yielded mice with abnormal primary follicles and germinal centres of the spleen and Peyer's patches, reflecting the inability of B lymphocytes to migrate into B-cell areas[47]. A novel peptide that binds and activates BRL-1 was recently discovered from the EST database[48,49].

In view of the number of novel GPCRs that have been cloned and are continuing to be discovered, it is expected that many endogenous ligands will be discovered. Unquestionably, this will result in an increase in the knowledge of the diversity in intercellular signalling mechanisms and should lead to novel insights into complex or poorly understood human disorders; it will also expand the boundaries of pharmacology. In conclusion, the discovery of the endogenous ligands will help determine the precise physiological role for each oGPCR. As the functions of these novel receptors are uncovered, they could become targets for the development of new pharmacological therapies for diseases not previously considered amenable to pharmacological therapy.

Selected references

1 Marchese, A., George, S. and O'Dowd, B. (1998) in *Identification and Expression of G Protein-coupled Receptors* (Lynch, K., ed.), pp. 1–26, John Wylie & Sons
2 Civelli, O. *et al.* (1997) *J. Recept. Signal. Transduct. Res.* 17, 545–550
3 Mollereau, C. *et al.* (1999) *Mol. Pharmacol.* 55, 324–331
4 Zondag, G. *et al.* (1998) *Biochem. J.* 330, 605–609
5 Okamoto, H. *et al.* (1998) *J. Biol. Chem.* 273, 27104–27110
6 Lee, M. J. *et al.* (1998) *Science* 279, 1552–1555
7 An, S. *et al.* (1998) *J. Biol. Chem.* 273, 7906–7910
8 Hecht, J. H. *et al.* (1996) *J. Cell Biol.* 135, 1071–1083
9 An, S. *et al.* (1998) *Mol. Pharmacol.* 54, 881–888
10 Maenhaut, C. *et al.* (1990) *Biochem. Biophys. Res. Commun.* 173, 1169–1178
11 Matsuda, L. A. *et al.* (1990) *Nature* 346, 561–564
12 Yoshida, R. *et al.* (1997) *J. Biol. Chem.* 272, 13803–13809
13 Hieshima, K. *et al.* (1997) *J. Biol. Chem.* 272, 5846–5853
14 Power, C. A. *et al.* (1997) *J. Exp. Med.* 186, 825–835
15 Liao, F., Lee, H. and Farber, J. (1997) *Genomics* 40, 175–180
16 Baba, M. *et al.* (1997) *J. Biol. Chem.* 272, 14893–14898
17 Wells, T. N. and Peitsch, M. C. (1997) *J. Leukocyte Biol.* 61, 545–550
18 Legler, D. F. *et al.* (1998) *J. Exp. Med.* 187, 655–660
19 Pan, Y. *et al.* (1997) *Nature* 387, 611–617
20 Bazan, J. F. *et al.* (1997) *Nature* 385, 640–644
21 Imai, T. *et al.* (1997) *Cell* 91, 521–530
22 Heiber, M. *et al.* (1995) *DNA Cell Biol.* 14, 25–35
23 Yoshida, T. *et al.* (1998) *J. Biol. Chem.* 273, 16551–16554
24 Stadel, J. M., Wilson, S. and Bergsma, D. J. (1997) *Trends Pharmacol. Sci.* 18, 430–437
25 Sakurai, T. *et al.* (1998) *Cell* 92, 573–585
26 de Lecea, L. *et al.* (1998) *Proc. Natl. Acad. Sci. U. S. A.* 95, 322–327
27 Hinuma, S. *et al.* (1998) *Nature* 393, 272–276
28 Marchese, A. *et al.* (1995) *Genomics* 29, 335–344
29 Tatemoto, K. *et al.* (1998) *Biochem. Biophys. Res. Commun.* 251, 471–476
30 O'Dowd, B. F. *et al.* (1993) *Gene* 136, 355–360
31 Klein, C. *et al.* (1998) *Nat. Biotechnol* 16, 1334–1337
32 Monnot, C. *et al.* (1991) *Mol. Endocrinol.* 5, 1477–1487
33 Barak, L. S., Ferguson, S. S. G., Zhang, J. and Caron, M. G. (1997) *J. Biol. Chem.* 272, 27497–27500
34 Graminski, G. F., Jayawickreme, C. K., Potenza, M. N. and Lerner, M. R. (1993) *J. Biol. Chem.* 268, 5957–5964
35 Potenza, M. N., Graminski, G. F. and Lerner, M. R. (1992) *Anal. Biochem.* 206, 315–322
36 McClintock, T. S. *et al.* (1993) *Anal. Biochem.* 209, 298–305
37 Milligan, G., Marshall, F. and Rees, S. (1996) *Trends Pharmacol. Sci.* 17, 235–237
38 Offermanns, S. and Simon, M. I. (1995) *J. Biol. Chem.* 270, 15175–15180
39 Arai, H. and Charo, I. F. (1996) *J. Biol. Chem.* 271, 21814–21819
40 Njuki, F. *et al.* (1993) *Clin. Sci.* 85, 385–388
41 McLatchie, L. *et al.* (1998) *Nature* 393, 333–339
42 Jones, K. A. *et al.* (1998) *Nature* 396, 674–679
43 White, J. H. *et al.* (1998) *Nature* 396, 679–682
44 Kaupmann, K. *et al.* (1998) *Nature* 396, 683–686
45 Ng, G. *et al.* (1999) *J. Biol. Chem.* 274, 7607–7610
46 Kuner, R. *et al.* (1999) *Science* 283, 74–77
47 Forster, R. *et al.* (1996) *Cell* 87, 1037–1047
48 Legler, D. F. *et al.* (1998) *J. Exp. Med.* 187, 655–660
49 Gunn, M. D. (1998) *Nature* 391, 799–803

Acknowledgements
The authors' research is supported by grants from the Medical Research Council of Canada, the National Institute for Drug Abuse, and the Smokeless Tobacco Research Council.

MOLECULAR CARCINOGENESIS 24:153–159 (1999)

IN PERSPECTIVE
Claudio J. Conti, Editor

Microarrays and Toxicology: The Advent of Toxicogenomics

Emile F. Nuwaysir,[1] Michael Bittner,[2] Jeffrey Trent,[2] J. Carl Barrett,[1] and Cynthia A. Afshari[1]

[1]*Laboratory of Molecular Carcinogenesis, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina*
[2]*Laboratory of Cancer Genetics, National Human Genome Research Institute, Bethesda, Maryland*

The availability of genome-scale DNA sequence information and reagents has radically altered life-science research. This revolution has led to the development of a new scientific subdiscipline derived from a combination of the fields of toxicology and genomics. This subdiscipline, termed toxicogenomics, is concerned with the identification of potential human and environmental toxicants, and their putative mechanisms of action, through the use of genomics resources. One such resource is DNA microarrays or "chips," which allow the monitoring of the expression levels of thousands of genes simultaneously. Here we propose a general method by which gene expression, as measured by cDNA microarrays, can be used as a highly sensitive and informative marker for toxicity. Our purpose is to acquaint the reader with the development and current state of microarray technology and to present our view of the usefulness of microarrays to the field of toxicology. *Mol. Carcinog. 24:153–159, 1999.*

Key words: toxicology; gene expression; animal bioassay

INTRODUCTION

Technological advancements combined with intensive DNA sequencing efforts have generated an enormous database of sequence information over the past decade. To date, more than 3 million sequences, totaling over 2.2 billion bases [1], are contained within the GenBank database, which includes the complete sequences of 19 different organisms [2]. The first complete sequence of a free-living organism, *Haemophilus influenzae,* was reported in 1995 [3] and was followed shortly thereafter by the first complete sequence of a eukaryote, *Saccharomyces cervisiae* [4]. The development of dramatically improved sequencing methodologies promises that complete elucidation of the *Homo sapiens* DNA sequence is not far behind [5].

To exploit more fully the wealth of new sequence information, it was necessary to develop novel methods for the high-throughput or parallel monitoring of gene expression. Established methods such as northern blotting, RNAse protection assays, S1 nuclease analysis, plaque hybridization, and slot blots do not provide sufficient throughput to effectively utilize the new genomics resources. Newer methods such as differential display [6], high-density filter hybridization [7,8], serial analysis of gene expression [9], and cDNA- and oligonucleotide-based microarray "chip" hybridization [10–12] are possible solutions to this bottleneck. It is our belief that the microarray approach, which allows the monitoring of expression levels of thousands of genes simultaneously, is a tool of unprecedented power for use in toxicology studies.

Almost without exception, gene expression is altered during toxicity, as either a direct or indirect result of toxicant exposure. The challenge facing toxicologists is to define, under a given set of experimental conditions, the characteristic and specific pattern of gene expression elicited by a given toxicant. Microarray technology offers an ideal platform for this type of analysis and could be the foundation for a fundamentally new approach to toxicology testing.

MICROARRAY DEVELOPMENT AND APPLICATIONS

cDNA Microarrays

In the past several years, numerous systems were developed for the construction of large-scale DNA arrays. All of these platforms are based on cDNAs or oligonucleotides immobilized to a solid support. In the cDNA approach, cDNA (or genomic) clones of interest are arrayed in a multi-well format and amplified by polymerase chain reaction. The products of this amplification, which are usually 500- to 2000-bp clones from the 3′ regions of the genes of interest, are then spotted onto solid support by using high-speed robotics. By using this method, microarrays of up to 10 000 clones can be generated by spotting onto a glass substrate

*Correspondence to: Laboratory of Molecular Carcinogenesis, National Institute of Environmental Health Sciences, 111 Alexander Drive, Research Triangle Park, NC 27709.

Received 8 December 1998; Accepted 5 January 1999

Abbreviations: PAH, polycyclic aromatic hydrocarbon; NIEHS, National Institute of Environmental Health Sciences.

[13,14]. Sample detection for microarrays on glass involves the use of probes labeled with fluorescent or radioactive nucleotides.

Fluorescent cDNA probes are generated from control and test RNA samples in single-round reverse-transcription reactions in the presence of fluorescently tagged dUTP (e.g., Cy3-dUTP and Cy5-dUTP), which produces control and test products labeled with different fluors. The cDNAs generated from these two populations, collectively termed the "probe," are then mixed and hybridized to the array under a glass coverslip [10,11,15]. The fluorescent signal is detected by using a custom-designed scanning confocal microscope equipped with a motorized stage and lasers for fluor excitation [10,11,15]. The data are analyzed with custom digital image analysis software that determines for each DNA feature the ratio of fluor 1 to fluor 2, corrected for local background [16,17]. The strength of this approach lies in the ability to label RNAs from control and treated samples with different fluorescent nucleotides, allowing for the simultaneous hybridization and detection of both populations on one microarray. This method eliminates the need to control for hybridization between arrays. The research groups of Drs. Patrick Brown and Ron Davis at Stanford University spearheaded the effort to develop this approach, which has been successfully applied to studies of *Arabidopsis thaliana* RNA [10], yeast genomic DNA [15], tumorigenic versus non-tumorigenic human tumor cell lines [11], human T-cells [18], yeast RNA [19], and human inflammatory disease–related genes [20]. The most dramatic result of this effort was the first published account of gene expression of an entire genome, that of the yeast *Saccharomyces cervisiae* [21].

In an alternative approach, large numbers of cDNA clones can be spotted onto a membrane support, albeit at a lower density [7,22]. This method is useful for expression profiling and large-scale screening and mapping of genomic or cDNA clones [7,22–24]. In expression profiling on filter membranes, two different membranes are used simultaneously for control and test RNA hybridizations, or a single membrane is stripped and reprobed. The signal is detected by using radioactive nucleotides and visualized by phosphorimager analysis or autoradiography. Numerous companies now sell such cDNA membranes and software to analyze the image data [25–27].

Oligonucleotide Microarrays

Oligonucleotide microarrays are constructed either by spotting prefabricated oligos on a glass support [13] or by the more elegant method of direct in situ oligo synthesis on the glass surface by photolithography [28–30]. The strength of this approach lies in its ability to discriminate DNA molecules based on single base-pair difference. This allows the application of this method to the fields of medical diagnostics, pharmacogenetics, and sequencing by hybridization as well as gene-expression analysis.

Fabrication of oligonucleotide chips by photolithography is theoretically simple but technically complex [29,30]. The light from a high-intensity mercury lamp is directed through a photolithographic mask onto the silica surface, resulting in deprotection of the terminal nucleotides in the illuminated regions. The entire chip is then reacted with the desired free nucleotide, resulting in selected chain elongation. This process requires only 4n cycles (where n = oligonucleotide length in bases) to synthesize a vast number of unique oligos, the total number of which is limited only by the complexity of the photolithographic mask and the chip size [29,31,32].

Sample preparation involves the generation of double-stranded cDNA from cellular poly(A)+ RNA followed by antisense RNA synthesis in an in vitro transcription reaction with biotinylated or fluor-tagged nucleotides. The RNA probe is then fragmented to facilitate hybridization. If the indirect visualization method is used, the chips are incubated with fluor-linked streptavidin (e.g., phycoerythrin) after hybridization [12,33]. The signal is detected with a custom confocal scanner [34]. This method has been applied successfully to the mapping of genomic library clones [35], to de novo sequencing by hybridization [28,36], and to evolutionary sequence comparison of the *BRCA1* gene [37]. In addition, mutations in the cystic fibrosis [38] and BRCA1 [39] gene products and polymorphisms in the human immunodeficiency virus-1 clade B protease gene [40] have been detected by this method. Oligonucleotide chips are also useful for expression monitoring [33] as has been demonstrated by the simultaneous evaluation of gene-expression patterns in nearly all open reading frames of the yeast strain *S. cerevisiae* [12]. More recently, oligonucleotide chips have been used to help identify single nucleotide polymorphisms in the human [41] and yeast [42] genomes.

THE USE OF MICROARRAYS IN TOXICOLOGY

Screening for Mechanism of Action

The field of toxicology uses numerous in vivo model systems, including the rat, mouse, and rabbit, to assess potential toxicity and these bioassays are the mainstay of toxicology testing. However, in the past several decades, a plethora of in vitro techniques have been developed to measure toxicity, many of which measure toxicant-induced DNA damage. Examples of these assays include the Ames test, the Syrian hamster embryo cell transformation assay, micronucleus assays, measurements of sister chromatid exchange and unscheduled DNA synthesis, and many others. Fundamental to all of these methods is the fact that toxicity is often preceded by, and results in, alterations in gene expression. In many cases, these changes in gene expression are a

far more sensitive, characteristic, and measurable endpoint than the toxicity itself. We therefore propose that a method based on measurements of the genome-wide gene expression pattern of an organism after toxicant exposure is fundamentally informative and complements the established methods described above.

We are developing a method by which toxicants can be identified and their putative mechanisms of action determined by using toxicant-induced gene expression profiles. In this method, in one or more defined model systems, dose and time-course parameters are established for a series of toxicants within a given prototypic class (e.g., polycyclic aromatic hydrocarbons (PAHs)). Cells are then treated with these agents at a fixed toxicity level (as measured by cell survival), RNA is harvested, and toxicant-induced gene expression changes are assessed by hybridization to a cDNA microarray chip (Figure 1). We have developed a custom DNA chip, called ToxChip v1.0, specifically for this purpose and will discuss it in more detail below. The changes in gene expression induced by the test agents in the model systems are analyzed, and the common set of changes unique to that class of toxicants, termed a toxicant signature, is determined.

This signature is derived by ranking across all experiments the gene-expression data based on relative fold induction or suppression of genes in treated samples versus untreated controls and selecting the most consistently different signals across the sample set. A different signature may be established for each prototypic toxicant class. Once the signatures are determined, gene-expression profiles induced by unknown agents in these same model systems can then be compared with the established signatures. A match assigns a putative mechanism of action to the test compound. Figure 2 illustrates this signature method for different types of oxidant stressors, PAHs, and peroxisome proliferators. In this example, the unknown compound in question had a gene-expression profile similar to that of the oxidant stressors in the database. We anticipate that this general method will also reveal cross talk between different pathways induced by a single agent (e.g., reveal that a compound has both PAH-like and oxidant-like properties). In the future, it may be necessary to distinguish very subtle differences between compounds within a very large sample set (e.g., thousands of highly similar structural isomers in a combinatorial chemistry library or peptide library). To generate these highly refined signatures, standard statistical clustering techniques or principal-component analysis can be used.

For the studies outlined in Figure 2, we developed the custom cDNA microarray chip ToxChip v1.0.

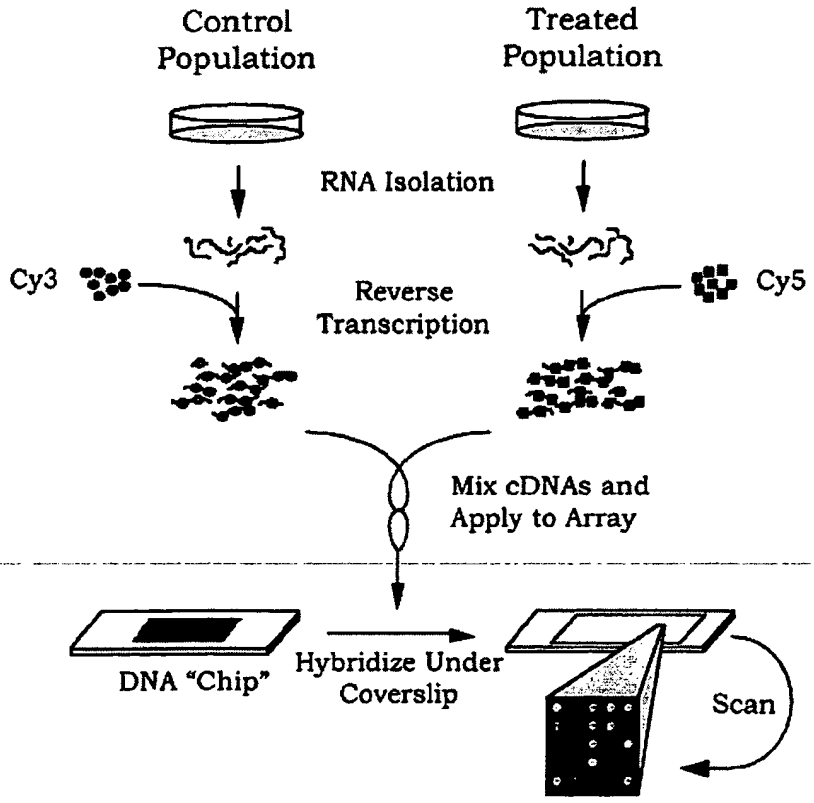

Figure 1. Simplified overview of the method for sample preparation and hybridization to cDNA microarrays. For illustrative purposes, samples derived from cell culture are depicted, although other sample types are amenable to this analysis.

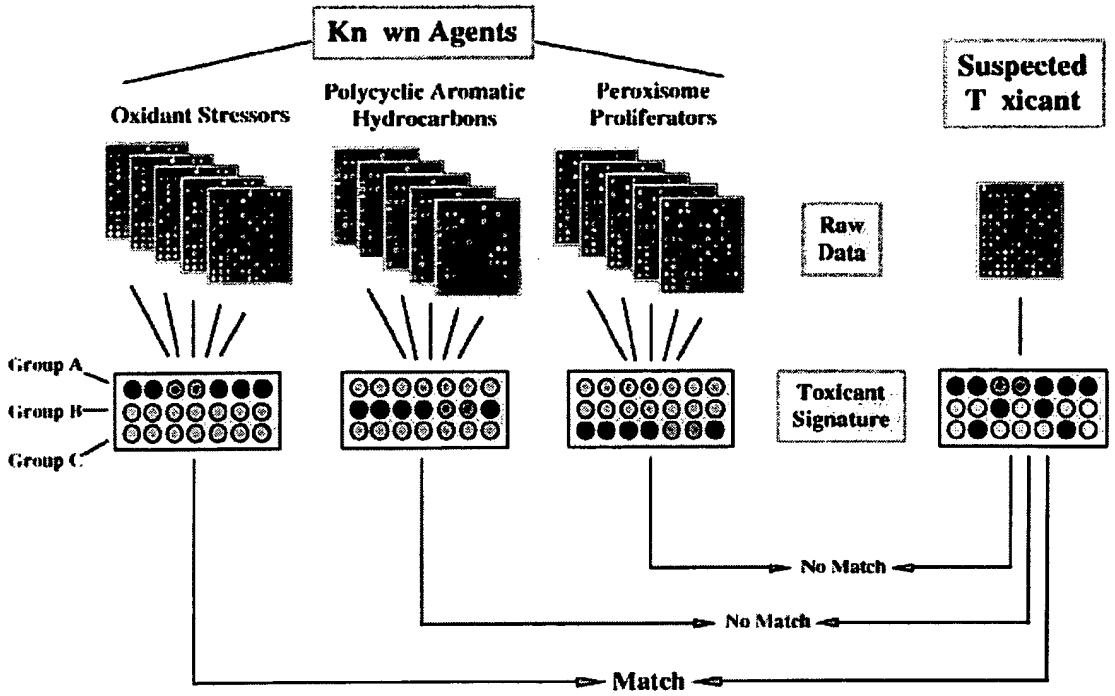

Figure 2. Schematic representation of the method for identification of a toxicant's mechanism of action. In this method, gene-expression data derived from exposure of model systems to known toxicants are analyzed, and a set of changes characteristic to that type of toxicant (termed the toxicant signature) is identified. As depicted, oxidant stressors produce consistent changes in group A genes (indicated by red and green circles), but not group B or C genes (indicated by gray circles). The set of gene-expression changes elicited by the suspected toxicant is then compared with these characteristic patterns, and a putative mechanism of action is assigned to the unknown agent.

The 2090 human genes that comprise this subarray were selected for their well-documented involvement in basic cellular processes as well as their responses to different types of toxic insult. Included on this list are DNA replication and repair genes, apoptosis genes, and genes responsive to PAHs and dioxin-like compounds, peroxisome proliferators, estrogenic compounds, and oxidant stress. Some of the other categories of genes include transcription factors, oncogenes, tumor suppressor genes, cyclins, kinases, phosphatases, cell adhesion and motility genes, and homeobox genes. Also included in this group are 84 housekeeping genes, whose hybridization intensity is averaged and used for signal normalization of the other genes on the chip. To date, very few toxicants have been shown to have appreciable effects on the expression of these housekeeping genes. However, this housekeeping list will be revised if new data warrant the addition or deletion of a particular gene. Table 1 contains a general description of some of the different classes of genes that comprise ToxChip v1.0.

When a toxicant signature is determined, the genes within this signature are flagged within the database. When uncharacterized toxicants are then screened, the data can be quickly reformatted so that blocks of genes representing the different signatures are displayed [11]. This facilitates rapid, visual interpretation of data. We are also developing ToxChip v2.0 and chips for other model systems, including rat, mouse, *Xenopus*, and yeast, for use in toxicology studies.

Animal Models in Toxicology Testing

The toxicology community relies heavily on the use of animals as model systems for toxicology testing. Unfortunately, these assays are inherently expensive, require large numbers of animals and take a long time to complete and analyze. Therefore, the National Institute of Environmental Health Sciences (NIEHS), the National Toxicology Program, and the toxicology community at large are committed to reducing the number of animals used, by developing more efficient and alternative testing methodologies. Although substantial progress has been made in the development of alternative methods, bioassays are still used for testing endpoints such as neurotoxicity, immunotoxicity, reproductive and developmental toxicology, and genetic toxicology. The rodent cancer bioassay is a particularly expensive and time-consuming assay, as it requires almost 4 yr, 1200 animals, and millions of dollars to execute and analyze [43]. In vitro experiments of the type outlined in Figure 2 might provide evidence that an unknown

Table 1. ToxChip v1.0: A Human cDNA Microarray Chip Designed to Detect Responses to Toxic Insult

Gene category	No. of genes on chip
Apoptosis	72
DNA replication and repair	99
Oxidative stress/redox homeostasis	90
Peroxisome proliferator responsive	22
Dioxin/PAH responsive	12
Estrogen responsive	63
Housekeeping	84
Oncogenes and tumor suppressor genes	76
Cell-cycle control	51
Transcription factors	131
Kinases	276
Phosphatases	88
Heat-shock proteins	23
Receptors	349
Cytochrome P450s	30

*This list is intended as a general guide. The gene categories are not unique, and some genes are listed in multiple categories.

agent is (or is not) responsible for eliciting a given biological response. This information would help to select a bioassay more specifically suited to the agent in question or perhaps suggest that a bioassay is not necessary, which would dramatically reduce cost, animal use, and time.

The addition of microarray techniques to standard bioassays may dramatically enhance the sensitivity and interpretability of the bioassay and possibly reduce its cost. Gene-expression signatures could be determined for various types of tissue-specific toxicants, and new compounds could be screened for these characteristic signatures, providing a rapid and sensitive in vivo test. Also, because gene expression is often exquisitely sensitive to low doses of a toxicant, the combination of gene-expression screening and the bioassay might allow the use of lower toxicant doses, which are more relevant to human exposure levels, and the use of fewer animals. In addition, gene-expression changes are normally measured in hours or days, not in the months to years required for tumor development. Furthermore, microarrays might be particularly useful for investigating the relationship between acute and chronic toxicity and identifying secondary effects of a given toxicant by studying the relationship between the duration of exposure to a toxicant and the gene-expression profile produced. Thus, a bioassay that incorporates gene-expression signatures with traditional endpoints might be substantially shorter, use more realistic dose regimens, and cost substantially less than the current assays do.

These considerations are also relevant for branches of toxicology not related to human health and not using rodents as model systems, such as aquatic toxicology and plant pathology. Bioassays based on the flathead minnow, *Daphnia,* and *Arabadopsis* could also be improved by the addition of microarray analysis. The combination of microarrays with traditional bioassays might also be useful for investigating some of the more intractable problems in toxicology research, such as the effects of complex mixtures and the difficulties in cross-species extrapolation.

Exposure Assessment, Environmental Monitoring, and Drug Safety

The currently used methods for assessment of exposure to chemical toxicants are based on measurement of tissue toxin levels or on surrogate markers of toxicity, termed biomarkers (e.g., peripheral blood levels of hepatic enzymes or DNA adducts). Because gene expression is a sensitive endpoint, gene expression as measured with microarray technology may be useful as a new biomarker to more precisely identify hazards and to assess exposure. Similarly, microarrays could be used in an environmental-monitoring capacity to measure the effect of potential contaminants on the gene-expression profiles of resident organisms. In an analogous fashion, microarrays could be used to measure gene-expression endpoints in subjects in clinical trials. The combination of these gene-expression data and more established toxic endpoints in these trials could be used to define highly precise surrogates of safety.

Gene-expression profiles in samples from exposed individuals could be compared to the profiles of the same individuals before exposure. From this information, the nature of the toxic exposure can be determined or a relative clinical safety factor estimated. In the future it may also be possible to estimate not only the nature but the dose of the toxicant for a given exposure, based on relative gene-expression levels. This general approach may be particularly appropriate for occupational-health applications, in which unexposed and exposed samples from the same individuals may be obtainable. For example, a pilot study of gene expression in peripheral-blood lymphocytes of Polish coke-oven workers exposed to PAHs (and many other compounds) is under consideration at the NIEHS. An important consideration for these types of studies is that gene expression can be affected by numerous factors, including diet, health, and personal habits. To reduce the effects of these confounding factors, it may be necessary to compare pools of control samples with pools of treated samples. In the future it may be possible to compare exposed sample sets to a national database of human-expression data, thus eliminating the need to provide an unexposed sample from the same individual. Efforts to develop such a national gene-expression database are currently under way [44,45]. However, this national database approach will require a better understanding of genome-wide gene expression across the highly diverse human population and of the effects of environmental factors on this expression.

Alleles, Oligo Arrays, and Toxicogenetics

Gene sequences vary between individuals, and this variability can be a causative factor in human diseases of environmental origin [46,47]. A new area of toxicology, termed toxicogenetics, was recently developed to study the relationship between genetic variability and toxicant susceptibility. This field is not the subject of this discussion, but it is worthwhile to note that the ability of oligonucleotide arrays to discriminate DNA molecules based on single base-pair differences makes these arrays uniquely useful for this type of analysis. Recent reports demonstrated the feasibility of this approach [41,42]. The NIEHS has initiated the Environmental Genome Project to identify common sequence polymorphisms in 200 genes thought to be involved in environmental diseases [48]. In a pilot study on the feasibility of this application to the Environmental Genome Project, oligonucleotide arrays will be used to resequence 20 candidate genes. This toxicogenetic approach promises to dramatically improve our understanding of interindividual variability in disease susceptibility.

FUTURE PRIORITIES

There are many issues that must be addressed before the full potential of microarrays in toxicology research can be realized. Among these are model system selection, dose selection, and the temporal nature of gene expression. In other words, in which species, at what dose, and at what time do we look for toxicant-induced gene expression? If human samples are analyzed, how variable is global gene expression between individuals, before and after toxicant exposure? What are the effects of age, diet, and other factors on this expression? Experience, in the form of large data sets of toxicant exposures, will answer these questions.

One of the most pressing issues for array scientists is the construction of a national public database (linked to the existing public databases) to serve as a repository for gene-expression data. This relational database must be made available for public use, and researchers must be encouraged to submit their expression data so that others may view and query the information. Researchers at the National Institutes of Health have made laudable progress in developing the first generation of such a database [44,45]. In addition, improved statistical methods for gene clustering and pattern recognition are needed to analyze the data in such a public database.

The proliferation of different platforms and methods for microarray hybridizations will improve sample handling and data collection and analysis and reduce costs. However, the variety of microarray methods available will create problems of data compatibility between platforms. In addition, the near-infinite variety of experimental conditions under which data will be collected by different laboratories will make large-scale data analysis extremely difficult. To help circumvent these future problems, a set of standards to be included on all platforms should be established. These standards would facilitate data entry into the national database and serve as reference points for cross-platform and inter-laboratory data analysis.

Many issues remain to be resolved, but it is clear that new molecular techniques such as microarray hybridization will have a dramatic impact on toxicology research. In the future, the information gathered from microarray-based hybridization experiments will form the basis for an improved method to assess the impact of chemicals on human and environmental health.

ACKNOWLEDGMENTS

The authors would like to thank Drs. Robert Maronpot, George Lucier, Scott Masten, Nigel Walker, Raymond Tennant, and Ms. Theodora Deverenux for critical review of this manuscript. EFN was supported in part by NIEHS Training Grant #ES07017-24.

REFERENCES

1. http://www.ncbi.nlm.nih.gov/Web/Genbank/index.html
2. http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html
3. Fleischmann RD, Adams MD, White O, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 1995;269:496–512.
4. Goffeau A, Barrell BG, Bussey H, et al. Life with 6000 genes. Science 1996;274:546, 563–567.
5. http://www.perkin-elmer.com/press/prc5448.html
6. Liang P, Pardee AB. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. Science 1992;257:967–971.
7. Pietu G, Alibert O, Guichard V, et al. Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cDNA array. Genome Res 1996;6:492–503.
8. Zhao ND, Hashida H, Takahashi N, Misumi Y, Sakaki Y. High-density cDNA filter analysis—A novel approach for large-scale, quantitative analysis of gene expression. Gene 1995;156:207–213.
9. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. Science 1995;270:484–487.
10. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene-expression patterns with a complementary DNA microarray. Science 1995;270:467–470.
11. DeRisi J, Penland L, Brown PO, et al. use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat Genet 1996;14:457–460.
12. Wodicka L, Dong HL, Mittmann M, Ho MH, Lockhart DJ. Genome-wide expression monitoring in Saccharomyces cerevisiae. Nat Biotechnol 1997;15:1359–1367.
13. Marshall A, Hodgson J. DNA chips: An array of possibilities. Nat Biotechnol 1998;16:27–31.
14. http://www.synteni.com
15. Shalon D, Smith SJ, Brown PO. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Res 1996;6:639–645.
16. Chen Y, Dougherty ER, Bittner ML. Ratio-based decisions and the quantitative analysis of cDNA microarray images. Biomedical Optics 1997;2:364–374.
17. Khan J, Simon R, Bittner M, et al. Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. Cancer Res 1998;58:5009–5013.
18. Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW. Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. Proc Natl Acad Sci USA 1996; 93:10614–10619.

19. Lashkari DA, DeRisi JL, McCusker JH, et al. Yeast microarrays for genome wide parallel genetic and gene expression analysis. Proc Natl Acad Sci USA 1997;94:13057–13062.
20. Heller RA, Schena M, Chai A, et al. Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. Proc Natl Acad Sci USA 1997;94:2150–2155.
21. DeRisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 1997;278:680–686.
22. Drmanac S, Stavropoulos NA, Labat I, et al. Gene-representing cDNA clusters defined by hybridization of 57,419 clones from infant brain libraries with short oligonucleotide probes. Genomics 1996;37:29–40.
23. Milosavljevic A, Savkovic S, Crkvenjakov R, et al. DNA sequence recognition by hybridization to short oligomers: Experimental verification of the method on the E. coli genome. Genomics 1996;37:77–86.
24. Drmanac S, Drmanac R. Processing of cDNA and genomic kilobase-size clones for massive screening, mapping and sequencing by hybridization. Biotechniques 1994;17:328–329, 332–336.
25. http://www.resgen.com/
26. http://www.genomesystems.com/
27. http://www.clontech.com/
28. Pease AC, Solas DA, Fodor SPA. Parallel synthesis of spatially addressable oligonucleotide probe matrices. Abstract. Abstracts of Papers of the American Chemical Society 1992;203:34.
29. Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SPA. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. Proc Natl Acad Sci USA 1994;91:5022–5026.
30. Fodor SPA, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. Light-directed, spatially addressable parallel chemical synthesis. Science 1991;251:767–773.
31. McGall G, Labadie J, Brock P, Wallraff G, Nguyen T, Hinsberg W. Light-directed synthesis of high-density oligonucleotide arrays using semiconductor photoresists. Proc Natl Acad Sci USA 1996;93:13555–13560.
32. Lipshutz RJ, Morris D, Chee M, et al. Using oligonucleotide probe arrays to access genetic diversity. Biotechniques 1995;19:442–447.
33. Lockhart DJ, Dong HL, Byrne MC, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 1996;14:1675–1680.
34. http://www.mdyn.com/
35. Sapolsky RJ, Lipshutz RJ. Mapping genomic library clones using oligonucleotide arrays. Genomics 1996;33:445–456.
36. Chee M, Yang R, Hubbell E, et al. Accessing genetic information with high-density DNA arrays. Science 1996;274:610–614.
37. Hacia JG, Makalowski W, Edgemon K, et al. Evolutionary sequence comparisons using high-density oligonucleotide arrays. Nat Genet 1998;18:155–158.
38. Cronin MT, Fucini RV, Kim SM, Masino RS, Wespi RM, Miyada CG. Cystic fibrosis mutation detection by hybridization to light-generated DNA probe arrays. Hum Mutat 1996;7:244–255.
39. Hacia JG, Brody LC, Chee MS, Fodor SPA, Collins FS. Detection of heterozygous mutations in BRCA1 using high density oligonucleotide arrays and two-colour fluorescence analysis. Nat Genet 1996;14:441–447.
40. Kozal MJ, Shah N, Shen NP, et al. Extensive polymorphisms observed in HIV-1 clade B protease gene using high-density oligonucleotide arrays. Nat Med 1996;2:753–759.
41. Wang DG, Fan JB, Siao CJ, et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 1998;280:1077–1082.
42. Winzeler EA, Richards DR, Conway AR, et al. Direct allelic variation scanning of the yeast genome. Science 1998;281:1194–1197.
43. Chhabra RS, Huff JE, Schwetz BS, Selkirk J. An overview of prechronic and chronic toxicity carcinogenicity experimental-study designs and criteria used by the National Toxicology Program. Environ Health Perspect 1990;86:313–321.
44. Ermolaeva O, Rastogi M, Pruitt KD, et al. Data management and analysis for gene expression arrays. Nat Genet 1998;20:19–23.
45. http://www.nhgri.nih.gov/DIR/LCG/15K/HTML/dbase.html
46. Samson M, Libert F, Doranz BJ, et al. Resistance to HIV-1 infection in Caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. Nature 1996;382:722–725.
47. Bell DA, Taylor JA, Paulson DF, Robertson CN, Mohler JL, Lucier GW. Genetic risk and carcinogen exposure—A common inherited defect of the carcinogen-metabolism gene glutathione-S-transferase M1 (Gstm1) that increases susceptibility to bladder cancer. J Natl Cancer Inst 1993;85:1159–1164.
48. http://www.niehs.nih.gov/envgenom/home.html

Application of DNA Arrays to Toxicology

John C. Rockett and David J. Dix

Reproductive Toxicology Division, National Health and Environmental Effects Research Laboratory, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, USA

DNA array technology makes it possible to rapidly genotype individuals or quantify the expression of thousands of genes on a single filter or glass slide, and holds enormous potential in toxicologic applications. This potential led to a U.S. Environmental Protection Agency-sponsored workshop titled "Application of Microarrays to Toxicology" on 7–8 January 1999 in Research Triangle Park, North Carolina. In addition to providing state-of-the-art information on the application of DNA or gene microarrays, the workshop catalyzed the formation of several collaborations, committees, and user's groups throughout the Research Triangle Park area and beyond. Potential application of microarrays to toxicologic research and risk assessment include genome-wide expression analyses to identify gene-expression networks and toxicant-specific signatures that can be used to define mode of action, for exposure assessment, and for environmental monitoring. Arrays may also prove useful for monitoring genetic variability and its relationship to toxicant susceptibility in human populations. *Key words*: DNA arrays, gene arrays, microarrays, toxicology. *Environ Health Perspect* 107:681–685 (1999). [Online 6 July 1999]
http://ehpnet1.niehs.nih.gov/docs/1999/107p681-685rockett/abstract.html

Decoding the genetic blueprint is a dream that offers manifold returns in terms of understanding how organisms develop and function in an often hostile environment. With the rapid advances in molecular biology over the last 30 years, the dream has come a step closer to reality. Molecular biologists now have the ability to elucidate the composition of any genome. Indeed, almost 20 genomes have already been sequenced and more than 60 are currently under way. Foremost among these is the Human Genome Mapping Project. However, the genomes of a number of commonly used laboratory species are also under intensive investigation, including yeast, *Arabidopsis*, maize, rice, zebra fish, mouse, rat, and dog. It is widely expected that the completion of such programs will facilitate the development of many powerful new techniques and approaches to diagnosing and treating genetically and environmentally induced diseases which afflict mankind. However, the vast amount of data being generated by genome mapping will require new high-throughput technologies to investigate the function of the millions of new genes that are being reported. Among the most widely heralded of the new functional genomics technologies are DNA arrays, which represent perhaps the most anticipated new molecular biology technique since polymerase chain reaction (PCR).

Arrays enable the study of literally thousands of genes in a single experiment. The potential importance of arrays is enormous and has been highlighted by the recent publication of an entire *Nature Genetics* supplement dedicated to the technology (*1*). Despite this huge surge of interest, DNA arrays are still little used and largely unproven, as demonstrated by the high ratio of review and press articles to actual data papers. Even so, the potential they offer has driven venture capitalists into a frenzy of investment and many new companies are springing up to claim a share of this rapidly developing market.

The U.S. Environmental Protection Agency (EPA) is interested in applying DNA array technology to ongoing toxicologic studies. To learn more about the current state of the technology, the Reproductive Toxicology Division (RTD) of the National Health and Environmental Effects Research Laboratory (NHEERL; Research Triangle Park, NC) hosted a workshop on "Application of Microarrays to Toxicology" on 7–8 January 1999 in Research Triangle Park, North Carolina. The workshop was organized by David Dix, Robert Kavlock, and John Rockett of the RTD/NHEERL. Twenty-two intramural and extramural scientists from government, academia, and industry shared information, data, and opinions on the current and future applications for this exciting new technology. The workshop had more than 150 attendees, including researchers, students, and administrators from the EPA, the National Institute of Environmental Health Sciences (NIEHS), and a number of other establishments from Research Triangle Park and beyond. Presentations ranged from the technology behind array production through the sharing of actual experimental data and projections on the future importance and applications of arrays. The information contained in the workshop presentations should provide aid and insight into arrays in general and their application to toxicology in particular.

Array Elements

In the context of molecular biology, the word "array" is normally used to refer to a series of DNA or protein elements firmly attached in a regular pattern to some kind of supportive medium. DNA array is often used interchangeably with gene array or microarray. Although not formally defined, microarray is generally used to describe the higher density arrays typically printed on glass chips. The DNA elements that make up DNA arrays can be oligonucleotides, partial gene sequences, or full-length cDNAs. Companies offering pre-made arrays that contain less than full-length clones normally use regions of the genes which are specific to that gene to prevent false positives arising through cross-hybridization. Sequence verification of cDNA clone identity is necessary because of errors in identifying specific clones from cDNA libraries and databases. Premade DNA arrays printed on membranes are currently or imminently available for human, mouse, and rat. In most cases they contain DNA sequences representing several thousand different sequence clusters or genes as delineated through the National Center for Biotechnology Information UniGene Project (*2*). Many of these different UniGene clusters (putative genes) are represented only by expressed sequence tags (ESTs).

Array Printing

Arrays are typically printed on one of two types of support matrix. Nylon membranes are used by most off-the-shelf array providers such as Clontech Laboratories, Inc. (Palo Alto, CA), Genome Systems, Inc. (St. Louis, MO), and Research Genetics, Inc. (Huntsville, AL). Microarrays such as those produced by Affymetrix, Inc. (Santa Clara, CA), Incyte Pharmaceuticals, Inc. (Palo Alto, CA), and many do-it-yourself (DIY) arraying groups use glass wafers or slides. Although standard microscope slides may be used, they must be preprepared to facilitate sticking of the DNA to the glass. Several different

Address correspondence to J. Rockett, Reproductive Toxicology Division (MD-72), National Health and Environmental Effects Research Laboratory, U.S. EPA, Research Triangle Park, NC 27711 USA. Telephone: (919) 541-2678. Fax: (919) 541-4017. E-mail: rockett.john@epa.gov

The authors thank R. Kavlock for envisioning the application of array technology to toxicology at the U.S. Environmental Protection Agency. We also thank T. Wall and B. Deitz for administrative assistance.

This document has been reviewed in accordance with EPA policy and approved for publication. Mention of companies, trade names, or products does not signify endorsement of such by the EPA.

Received 23 March 1999; accepted 22 April 1999.

coatings have been successfully used, including silane and lysine. The coating of slides can easily be carried out in the laboratory, but many prefer the convenience of precoated slides available from suppliers.

Once the support matrix has been prepared, the DNA elements can be applied by several methods. Affymetrix, Inc., has developed a unique photolithographic technology for attaching oligonucleotides to glass wafers. More commonly, DNA is applied by either noncontact or contact printing. Noncontact printers can use thermal, solenoid, or piezoelectric technology to spray aliquots of solution onto the support matrix and may be used to produce slide or membrane-based arrays. Cartesian Technologies, Inc. (Irvine, CA) has developed nQUAD technology for use in its PixSys printers. The system couples a syringe pump with the microsolenoid valve, a combination that provides rapid quantitative dispensing of nanoliter volumes (down to 4.2 nL) over a variable volume range. A different approach to noncontact printing uses a solid pin and ring combination (Genetic MicroSystems, Inc., Woburn, MA). This system (Figure 1) allows a broader range of sample, including cell suspensions and particulates, because the printing head cannot be blocked up in the same way as a spray nozzle. Fluid transfer is controlled in this system primarily by the pin dimensions and the force of deposition, although the nature of the support matrix and the sample will also affect transfer to some degree.

In contact printing, the pin head is dipped in the sample and then touched to the support matrix to deposit a small aliquot. Split pins were one of the first contact-printing devices to be reported and are the suggested format for DIY arrayers, as described by Brown (*3*). Split pins are small metal pins with a precise groove cut vertically in the middle of the pin tip. In this system, 1–48 split pins are positioned in the pin-head. The split pins work by simple capillary action, not unlike a fountain pen—when the pin heads are dipped in the sample, liquid is drawn into the pin groove. A small (fixed) volume is then deposited each time the split pins are gently touched to the support matrix. Sample (100–500 pL depending on a variety of parameters) can be deposited on multiple slides before refilling is required, and array densities of > 2,500 spots/cm^2 may be produced. The deposit volume depends on the split size, sample fluidity, and the speed of printing. Split pins are relatively simple to produce and can be made in-house if a suitable machine shop is available. Alternatively, they can be obtained directly from companies such as TeleChem International, Inc. (Sunnyvale, CA).

Irrespective of their source, printers should be run through a preprint sequence prior to producing the actual experimental arrays; the first 100 or so spots of a new run tend to be somewhat variable. Factors effecting spot reproducibility include slide treatment homogeneity, sample differences, and instrument errors. Other factors that come into play include clean ejection of the drop and clogging (nQUAD printing) and mechanical variations and long-term alteration in print-head surface of solid and split pins. However, with careful preparation it is possible to get a coefficient of variance for spot reproducibility below 10%.

One potential printing problem is sample carryover. Repeated washing, blotting, and drying (vacuum) of print pins between samples is normally effective at reducing sample carryover to negligible amounts. Printing should also be carried out in a controlled environment. Humidified chambers are available in which to place printers. These help prevent dust contamination and produce a uniform drying rate, which is important in determining spot size, quality, and reproducibility.

In summary, although several printing technologies are available, none are particularly outstanding and the bottom line is that they are still in a relatively early stage of evolution.

Array Hybridization

The hybridization protocol is, practically speaking, relatively straightforward and those with previous experience in blotting should have little difficulty. Array hybridizations are, in essence, reverse Southern/Northern blots—instead of applying a labeled probe to the target population of DNA/RNA, the labeled population is applied to the probe(s). With membrane-based arrays, the control and treated mRNA populations are normally converted to cDNA and labeled with isotope (e.g., ^{33}P) in the process. These labeled populations are then hybridized independently to parallel or serial arrays and the hybridization signal is detected with a phosporimager. A less commonly used alternative to radioactive probes is enzymatic detection. The probe may be biotinylated, haptenylated, or have alkaline phosphatase/horseradish peroxidase attached. Hybridization is detected by enzymatic reaction yielding a color reaction (*4*). Differences in hybridization signals can be detected by eye or, more accurately, with the help of digital imaging and commercially available software. The labeling of the test populations for slide-based microarrays uses a slightly different approach. The probe typically consists of two samples of polyA$^+$ RNA (usually from a treated and a control population) that are converted to cDNA; in the process each is labeled with a different fluor. The independently labeled probes are then mixed together and hybridized to a single microarray slide and the resulting combined fluorescent signal is scanned. After normalization, it is possible to determine the ratio of fluorescent signals from a single hybridization of a slide-based microarray.

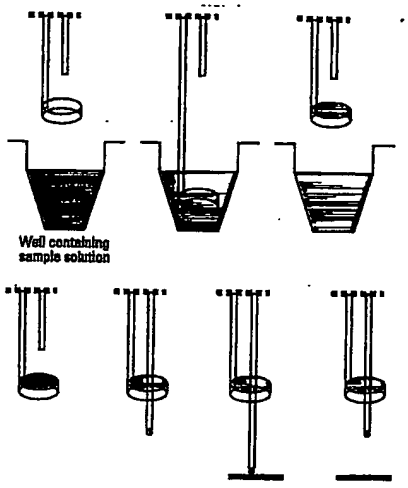

Figure 1. Genetic Microsystems (Woburn, MA) pin ring system for printing arrays. The pin ring combination consists of a circular open ring oriented parallel to the sample solution, with a vertical pin centered over the ring. When the ring is dipped into a solution and lifted, it withdraws an aliquot of sample held by surface tension. To spot the sample, the pin is driven down through the ring and a portion of the solution is transferred to the bottom of the pin. The pin continues to move downward until the pendant drop of solution makes contact with the underlying surface. The pin is then lifted, and gravity and surface tension cause deposition of the spot onto the array. Figure from Flowers et al. (*14*), with permission from Genetic Microsystems.

cDNA derived from control and treated populations of RNA is most commonly hybridized to arrays, although subtractive hybridization or differential display reactions may also be used. Fluorophore- or radiolabeled nucleotides are directly incorporated into the cDNA in the process of converting RNA to cDNA. Alternatively, 5′ end-labeled primers may be used for cDNA synthesis. These are labeled with a fluorophore for direct visualization of the hybridized array. Alternatively, biotin or a hapten may be attached to the primer, in which case fluor-labeled streptavidin or antibody must be applied before a signal can be generated. The most commonly used fluorophores at present are cyanine (Cy)3 and Cy5 (Amersham Pharmacia Biotech AB, Uppsala, Sweden). However, the relative expense of these fluorescent conjugates has driven a search for cheaper alternatives. Fluorescein, rhodamine, and Texas red have all been used, and companies such as Molecular Probes, Inc. (Eugene, OR) are developing a series of labeled nucleotides with a wide range of excitation and emission spectra which may prove to function as well as the Cy dyes.

Analysis of DNA Microarrays

Membrane-based arrays are normally analyzed on film or with a phosphorimager, whereas chip-based arrays require more specialized scanning devices. These can be divided into three main groups: the charge-coupled device camera systems, the nonconfocal laser scanners, and the confocal laser scanners. The advantages and disadvantages of each system are listed in Table 1.

Because a typical spot on a microarray can contain > 10^8 molecules, it is clear that a large variation in signal strength may occur. Current scanners cannot work across this many orders of magnitude (4 or 5 is more typical). However, the scanning parameters can normally be adjusted to collect more or less signal, such that two or three scans of the same array should permit the detection of rare and abundant genes.

When a microarray is scanned, the fluorescent images are captured by software normally included with the scanner. Several commercial suppliers provide additional software for quantifying array images, but the software tools are constantly evolving to meet the developing needs of researchers, and it is prudent to define one's own needs and clarify the exact capabilities of the software before its purchase. Issues that should be considered include the following:

- Can the software locate offset spots?
- Can it quantitate across irregular hybridization signals?
- Can the arrayed genes be programmed in for easy identification and location?
- Can the software connect via the Internet to databases containing further information on the gene(s) of interest?

One of the key issues raised at the workshop was the sensitivity of microarray technology. Experiments by General Scanning, Inc. (Watertown, MA), have shown that by using the Cy dyes and their scanner, signal can be detected down to levels of < 1 fluor molecule per square micrometer, which translates to detecting a rare message at approximately one copy per cell or less.

Array Applications

Although arrays are an emerging technology certain to undergo improvement and alteration, they have already been applied usefully to a number of model systems. Arrays are at their most powerful when they contain the entire genome of the species they are being used to study. For this reason, they have strong support among researchers utilizing yeast and *Caenorhabditis elegans* (*5*). The genomes of both of these species have been sequenced and, in the case of yeast, deposited onto arrays for examination of gene expression (*6,7*). With both of these species, it is relatively easy to perturb individual gene expression. Indeed, *C. elegans* knockouts can be made simply by soaking the worms in an antisense solution of the gene to be knocked out.

Table 1. Advantages and disadvantages of different microarray scanning systems.

	CCD camera system	Nonconfocal laser scanner	Confocal laser scanner
Advantages	Few moving parts	Relatively simple optics	Small depth of focus reduces artifacts
	Fast scanning of bright samples	—	May have high light collection efficiency
Disadvantages	Less appropriate for dim samples	Low light collection efficiency	Small depth of focus requires scanning precision
	Optical scatter can limit performance	Background artifacts not rejected	
		Resolution typically low	

CCD, charge-coupled device.
From Kawasaki (*15*).

By a process of systematic gene disruption, it is now possible to examine the cause and effect relationships between different genes in these simple organisms. This kind of approach should help elucidate biochemical pathways and genetic control processes, deconvolute polygenic interactions, and define the architecture of the cellular network. A simple case study of how this can be achieved was presented by Butow [University of Texas Southwestern Medical Center, Dallas, TX (Figure 2)]. Although it is the phenotypic result of a single gene knockout that is being examined, the effect of such perturbation will almost always be polygenic. Polygenic interactions will become increasingly important as researchers begin to move away from single gene systems when examining the nature of toxicologic responses to external stimuli. This is especially important in toxicology because the phenotype produced by a given environmental insult is never the result of the action of a single gene; rather, it is a complex interaction of one or multiple cellular pathways. Phenomena such as quantitative trait (the continuous variation of phenotype), epistasis (the effect of alleles of one or more genes on the expression of other genes), and penetrance (proportion of individuals of a given genotype that display a particular phenotype) will become increasingly evident and important as toxicologists push toward the ultimate goal of matching the responses of individuals to different environmental stimuli.

Analysis of the transcriptome (the expression level of all the genes in a given cell population) was a use of arrays addressed by several speakers. Unfortunately, current gene nomenclature is often confusing in that single genes are allocated multiple names (usually as a result of independent discovery by different laboratories), and there was a call for standardization of gene nomenclature. Nevertheless, once a transcriptome has been assembled it can then be transferred onto arrays and used to screen any chosen system. The EPA MicroArray Consortium (EPAMAC) is assembling testes transcriptomes for human, rat, and mouse. In a slightly different approach, Nuwaysir et al. (*8*) describes how the NIEHS assembled what is effectively a "toxicological transcriptome"—a library of human and mouse genes that have previously been proven or implicated in responses to toxicologic insults. Clontech Laboratories, Inc. (Palo Alto, CA), has begun a similar process by developing stress/toxicology filter arrays of rat, mouse, and human genes. Thus, rather than being tissue or cell specific, these stress/toxicology arrays can be used across a variety of model systems to look for alterations in the expression of toxicologically important genes and define the new field of toxicogenomics. The potential to identify toxicant families based on tissue- or cell-specific gene expression could revolutionize drug testing. These molecular signatures or fingerprints could not only point to the possible toxicity/carcinogenicity of newly discovered compounds (Figure 3), but also aid in elucidating their mechanism of action through identification of gene expression networks. By extension, such signatures could provide easily identifiable biomarkers to assess the degree, time, and nature of exposure.

DNA arrays are primarily a tool for examining differential gene expression in a given model. In this context they are referred to as closed systems because they lack the ability of other differential expression technologies, e.g., differential display and subtractive hybridization, to detect previously unknown genes not present on the array. This would appear to limit the power of DNA arrays to the imaginations and preconceptions of the researcher in selecting genes previously characterized and thought to be involved in the model system. However, the various genome sequencing projects have created a new category of sequence—the EST—that has partially mollified this deficiency. ESTs are cDNAs expressed in a given tissue that, although they may share some degree of sequence similarity to previously characterized genes, have not been assigned specific genetic identity. By incorporating EST clones into an array, it is possible to monitor the expression of these unknown genes. This can enable the identification of previously uncharacterized genes that may have biologic

significance in the model system. Filter arrays from Research Genetics and slide arrays from Incyte Pharmaceuticals both incorporate large numbers of ESTs from a variety of species.

A further use of microarrays is the identification of single nucleotide polymorphisms (SNPs). These genomic variations are abundant—they occur approximately every 1 kb or so—and are the basis of restriction fragment length polymorphism analysis used in forensic analysis. Affymetrix, Inc., designed chips that contain multiple repeats of the same gene sequence. Each position is present with all four possible bases. After the hybridization of the sample, the degree of hybridization to the different sequences can be measured and the exact sequence of the target gene deduced. SNPs are thought to be of vital importance in drug metabolism and toxicology. For example, single base differences in the regulatory region or active site of some genes can account for huge differences in the activity of that gene. Such SNPs are thought to explain why some people are able to metabolize certain xenobiotics better than others. Thus, arrays provide a further tool for the toxicologist investigating the nature of susceptible subpopulations and toxicologic response.

There are still many wrinkles to be ironed out before arrays become a standard tool for toxicologists. The main issues raised at the workshop by those with hands-on experience were the following:

- Expense: the cost of purchasing/contracting this technology is still too great for many individual laboratories.
- Clones: the logistics of identifying, obtaining, and maintaining a set of nonredundant, noncontaminated, sequence-verified, species/cell/tissue/field-specific clones.
- Use of inbred strains: where whole-organism models are being used, the use of inbred strains is important to reduce the potentially confusing effects of the individual variation typically seen in outbred populations.
- Probe: the need for relatively large amounts of RNA, which limits the type of sample (e.g., biopsy) that can be used. Also, different RNA extraction methods can give different results.
- Specificity: the ability to discriminate accurately between closely related genes (e.g., the cytochrome p450 family) and splice variants.
- Quantitation: the quantitation of gene expression using gene arrays is still open to debate. One reason for this is the different incorporation of the labeling dyes. However, the main difficulty lies in knowing what to normalize against. One option is to include a large number of so-called housekeeping genes in the array. However, the expression of these genes often change depending on the tissue and the toxicant, so it is necessary to characterize the expression of these genes in the model system before utilizing them. This is clearly not a viable option when screening multiple new compounds. A second option is to include on the array genes from a nonrelated species (e.g., a plant gene on an animal array) and to spike the probe with synthetic RNA(s) complementary to the gene(s).
- Reproducibility: this is sometimes questionable, and a figure of approximately two or three repeats was used as the minimum number required to confirm initial findings. Again, however, most people advocated the use of Northern blots or reverse transcriptase PCR to confirm findings.
- Sensitivity: concerns were voiced about the number of target molecules that must be present in a sample for them to be detected on the array.
- Efficiency: reproducible identification of 1.5- to 2-fold differences in expression was reported, although the number of genes that undergo this level of change and remain undetected is open to debate. It is important that this level of detection be ultimately achieved because it is commonly perceived that some important transcription factors and their regulators respond at such low levels. In most cases, 3- to 5-fold was the minimum change that most were happy to accept.
- Bioinformatics: perhaps the greatest concern was how to accurately interpret the data with the greatest accuracy and efficiency. The biggest headache is trying to identify networks of gene expression that are common to different treatments or doses. The amount of data from a single experiment is huge. It may be that, in the future, several groups individually equipped with specialized software algorithms for studying their favorite genes or gene systems will be able to share the same hybridized chips. Thus, arrays could usher in a new perspective on collaboration and the sharing of data.

EPAMAC

Perhaps the main reason most scientists are unable to use array technology is the high cost involved, whether buying off-the-shelf membranes, using contract printing services, or

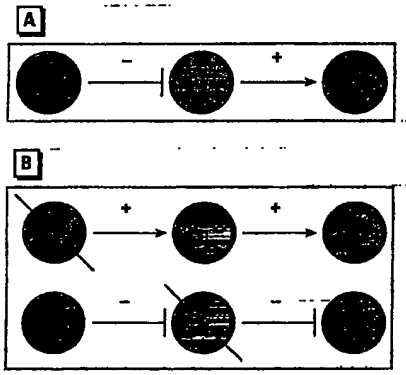

Figure 2. Potential effects of gene knockout within positively and negatively regulated gene expression networks. i_1 is limiting in wild type for expression of i_2. (A) A simple, two-component, linear regulatory network operating on gene i_2, where i_1 is a positive effector of i_2 and j_n is either a positive or negative effector of i_1. This network could be deduced by examining the consequence of (B) deleting j_n on the expression of i_1 and i_2, where the expression of i_2 would be decreased or increased depending on whether j_n was a positive or negative regulator. These and other connected components of even greater complexity could be revealed by genome-wide expression analysis. From Butow (15).

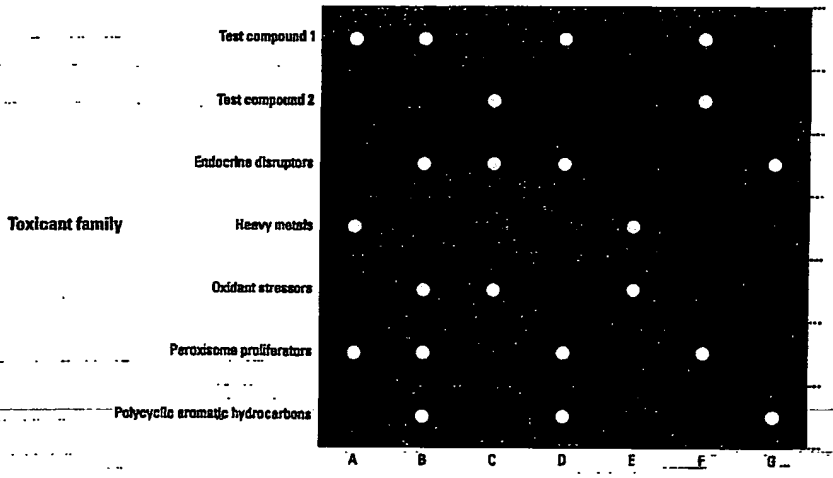

Figure 3. Gene expression profiles—also called fingerprints or signatures—of known toxicants or toxicant families may, in the future, be used to identify the potential toxicity of new drugs, etc. In this example, the genetic signature of test compound 1 is identical to that of known peroxisome proliferators, whereas that of test compound 2 does not match any known toxicant family. Based on these results, test compound 2 would be retained for further testing and test compound 1 would be eliminated.

producing chips in-house. In view of this, researchers at the RTD/NHEERL initiated the EPAMAC. This consortium brings together scientists from the EPA and a number of extramural labs with the aim of developing microarray capability through the sharing of resources and data. EPAMAC researchers are primarily interested in the developmental and toxicologic changes seen in testicular and breast tissue, and a portion of the workshop was set aside for EPAMAC members to share their ideas on how the experimental application of microarrays could facilitate their research. One of the central areas of interest to EPAMAC members is the effect of xenobiotics on male fertility and reproductive health. Of greatest concern is the effect of exposure during critical periods of development and germ cell differentiation (*9*), and how this may compromise sperm counts and quality following sexual maturation (*10*). As well as spermatogenic tissue, there is also interest in how residual mRNA found in mature sperm (*11*) could be used as an indicator of previous xenobiotic effects (it is easier to obtain a semen sample than a testicular biopsy). Arrays will be used to examine and compare the effect of exposure to heat and chemicals in testicular and epididymal gene expression profiles, with the aim of establishing relationships/associations between changes in developmental landmarks and the effects on sperm count and quality. Cluster, pattern, and other analysis of such data should help identify hidden relationships between genes that may reveal potential mechanisms of action and uncover roles for genes with unknown functions.

Summary

The full impact of DNA arrays may not be seen for several years, but the interest shown at this regional workshop indicates the high level of interest that they foster. Apart from educating and advertising the various technologies in this field, this workshop brought together a number of researchers from the Research Triangle Park area who are already using DNA arrays. The interest in sharing ideas and experiences led to the initiation of a Triangle array user's group.

Array technology is still in its infancy. This means that the hardware is still improving and there is no current consensus for standard procedures, quantitation, and interpretation. Consistency in spotting and scanning arrays is not yet optimized, and this is one of the most critical requirements of any experiment. In addition, one of the dark regions of array technology—strife in the courts over who owns what portions of it—has further muddled the future and is a potential barrier toward the development of consensus procedures.

Perhaps the greatest hurdle for the application of arrays is the actual interpretation of data. No specialists in bioinformatics attended the workshop, largely because they are rare and because as yet no one seems clear on the best method of approaching data analysis and interpretation. Cross-referencing results from multiple experiments (time, dose, repeats, different animals, different species) to identify commonly expressed genes is a great challenge. In most cases, we are still a long way from understanding how the expression of gene *X* is related to the expression of gene *Y*, and ordering gene expression to delineate causal relationships.

To the ordinary scientist in the typical laboratory, however, the most immediate problem is a lack of affordable instrumentation. One can purchase premade membranes at relatively affordable prices. Although these may be useful in identifying individual genes to pursue in more detail using other methods, the numbers that would be required for even a small routine toxicology experiment prohibit this as a truly viable approach. For the toxicologist, there is a need to carry out multiple experiments—dose responses, time curves, multiple animals, and repeats. Glass-based DNA arrays are most attractive in this context because they can be prepared in large batches from the same DNA source and accommodate control and treated samples on the same chip. Another problem with current off-the-shelf arrays is that they often do not contain one or more of the particular genes a group is interested in. One alternative is to obtain and/or produce a set of custom clones and have contract printing of membranes or slides carried out by a company such as Genomic Solutions, Inc. (Ann Arbor, MI). This approach is less expensive than laying out capital for one's own entire system, although at some point it might make economic sense to print one's own arrays.

Finally, DNA arrays are currently a team effort. They are a technology that uses a wide range of skills including engineering, statistics, molecular biology, chemistry, and bioinformatics. Because most individuals are skilled in only one or perhaps two of these areas, it appears that success with arrays may be best expected by teams of collaborators consisting of individuals having each of these skills.

Those considering array applications may be amused or goaded on by the following quote from *Fortune* magazine (*12*):

> Microprocessors have reshaped our economy, spawned vast fortunes and changed the way we live. Gene chips could be even bigger.

Although this comment may have been designed to excite the imagination rather than accurately reflect the truth, it is fair to say that the age of functional genomics is upon us. DNA arrays look set to be an important tool in this new age of biotechnology and will likely contribute answers to some of toxicology's most fundamental questions.

SPEAKERS

Cindy Afshari, NIEHS
Linda Birnbaum, U.S. EPA
Ron Butow, University of Texas Southwestern Medical Center
Alex Chenchik, Clontech Laboratories, Inc.
David Dix, U.S. EPA
Abdel Elkahloun, Research Genetics, Inc.
Sue Fenton, U.S. EPA
Norman Hecht, University of Pennsylvania
Pat Hurban, Paradigm Genetics, Inc.
Bob Kavlock, U.S. EPA
Ernie Kawasaki, General Scanning, Inc.
Steve Krawetz, Wayne State University
Nick Mace, Genetic Microsystems, Inc.
Scott Mordecai, Affymetrix, Inc.
Kevin Morgan, Glaxo Wellcome, Inc.
Elaine Poplin, Research Genetics, Inc.
Dod Rose, Cartesian Technologies, Inc.
Jim Samet, U.S. EPA
Sam Ward, University of Arizona
Jeff Welch, U.S. EPA
Reen Wu, University of California at Davis
Tim Zacharewski, Michigan State University

REFERENCES AND NOTES

1. The chipping forecast. Nat Genet 21(Suppl 1):3–60 (1999).
2. National Center for Biotechnology Information. The Unigene System. Available: www.ncbi.nlm.nih.gov/Schuler/UniGene [cited 22 March 1999].
3. Brown PO. The Brown Lab. Available: http://cmgm.Stanford.edu/pbrown [cited 22 March 1999].
4. Chen JJ, Wu R, Yang PC, Huang JY, Sher YP, Han MH, Kao WC, Lee PJ, Chiu TF, Chang F, et al. Profiling expression patterns and isolating differentially expressed genes by cDNA microarray system with colorimetry detection. Genomics 51:313–324 (1998).
5. Ward S. DNA Microarray Technology to Identify Genes Controlling Spermatogenesis. Available: www.mcb.arizona.edu/wardlab/microarray.html [cited 22 March 1999].
6. Marton MJ, DeRisi JL, Bennett HA, Iyer VR, Meyer MR, Roberts CJ, Stoughton R, Burchard J, Slade D, Dai H, et al. Drug target validation and identification of secondary drug target effects using DNA microarrays. Nat Med 4:1293–1301 (1998).
7. Brown PO. The Full Yeast Genome on a Chip. Available: http://cmgm.stanford.edu/pbrown/yeastchip.html [cited 22 March 1999].
8. Nuwaysir EF, Bittner M, Trent J, Barrett JC, Afshari CA. Microarrays and toxicology: the advent of toxicogenomics. Mol Carcinog 24(3):153–159 (1999).
9. Hecht NB. Molecular mechanisms of male germ cell differentiation. Bioessays 20:555–561 (1998).
10. Zacharewski TR. Timothy R. Zacharewski. Available: www.bch.msu.edu/faculty/zachar.htm [cited 22 March 1999].
11. Kramer JA, Krawetz SA. RNA in spermatozoa: implications for the alternative haploid genome. Mol Hum Reprod 3:473–478 (1997).
12. Stipp D. Gene chip breakthrough. Fortune, March 31:58–73 (1997).
13. Kawasaki E (General Scanning Instruments, Inc., Watertown, MA). Unpublished data.
14. Flowers P, Overbeck J, Mace ML Jr, Pegliughi FM, Eggers WJE, Yonkers H, Honkanen P, Montagu J, Rose SD. Development and Performance of a Novel Microarraying System Based on Surface Tension Forces. Available: http://www.geneticmicro.com/resources/html/coldspring.html [cited 22 March 1999].
15. Butow R (University of Texas Medical Center, Dallas, TX). Unpublished data.